العنوان:	التنبؤ بالمبيعات بإستخدام شجرة القرار: بالتطبيق على شركة سويتش للتجارة والهندسة قسم المولدات
المؤلف الرئيسي:	عبدالله، سمية عبداالله محمد الأمين
مؤلفين آخرين:	أحمد، شاذلى صديق محمد(مشرف)
التاريخ الميلادي:	2017
موقع:	الخرطوم
الصفحات:	1 - 82
رقم MD:	916054
نوع المحتوى:	رسائل جامعية
اللغة:	Arabic
الدرجة العلمية:	رسالة ماجستير
الجامعة:	جامعة النيلين
الكلية:	كلية علوم الحاسوب وتقانة المعلومات
الدولة:	السودان
قواعد المعلومات:	Dissertations
مواضيع:	التنبؤ بالمبيعات، شركة سويتش للتجارة والهندسة، قسم المولدات، تقنية أشجار القرار، التنقيب عن البيانات، قواعد البيانات
رابط:	http://search.mandumah.com/Record/916054

© 2024 المنظومة. جميع الحقوق محفوظة.
هذه المادة متاحة بناء على الإتفاق الموقع مع أصحاب حقوق النشر، علما أن جميع حقوق النشر محفوظة. يمكنك تحميل أو طباعة هذه المادة للاستخدام الشخصي فقط، ويمنع النسخ أو التحويل أو النشر عبر أي وسيلة (مثل مواقع الانترنت أو البريد الالكتروني) دون تصريح خطي من أصحاب حقوق النشر أو المنظومة.

للإستشهاد بهذا البحث قم بنسخ البيانات التالية حسب إسلوب الإستشهاد المطلوب:

إسلوب APA
عبداالله، سمية عبداالله محمد الأمين، و أحمد، شاذلى صديق محمد. (2017).التنبؤ بالمبيعات بإستخدام شجرة القرار: بالتطبيق على شركة سويتش للتجارة والهندسة قسم المولدات(رسالة ماجستير غير منشورة). جامعة النيلين، الخرطوم. مسترجع من http://search.mandumah.com/Record/916054

إسلوب MLA
عبداالله، سمية عبداالله محمد الأمين، و شاذلى صديق محمد أحمد. "التنبؤ بالمبيعات بإستخدام شجرة القرار: بالتطبيق على شركة سويتش للتجارة والهندسة قسم المولدات" رسالة ماجستير. جامعة النيلين، الخرطوم، 2017. مسترجع من http://search.mandumah.com/Record/916054

© 2024 المنظومة. جميع الحقوق محفوظة.
هذه المادة متاحة بناء على الإتفاق الموقع مع أصحاب حقوق النشر، علما أن جميع حقوق النشر محفوظة. يمكنك تحميل أو طباعة هذه المادة للاستخدام الشخصي فقط، ويمنع النسخ أو التحويل أو النشر عبر أي وسيلة (مثل مواقع الانترنت أو البريد الالكتروني) دون تصريح خطي من أصحاب حقوق النشر أو المنظومة.

الفصل الثاني

(الاطار النظري)

2-1الدراسات السابقة

2-1-1الدراسة الاولى بواسطة الطالبة / داليا عبد الحسين احمد دولة العراق بحث لنيل درجة الماجستير من م.م /الجامعة التكنولوجية بعنوان **دور تنقيب البيانات في زيادة أداءالمنظمة**(دراسة تحليليه في المصرف الصناعي ببغداد) إشراف أ .م .د .زكريا مطلك الدوري جامعة بغداد /كلية الادارة والاقتصاد المجلد/13ع/ 48 لسنة2007 .

تتلخص مشكلة البحث في ثلاث نقاط أساسية وجود عدد كبير من المقترضين في المصرف الصناعي ,وجود كم هائل من البيانات عنهم وعدم وجود تقنية تمكن المصرف من اتخاذ قرار سريع.

كما يمكن صياغة بعض التساؤلات مثل:ما مدى اهتمام المصرف في التوثيق الكامل للبيانات؟,كيف تستخدم الأساليب العلمية والتكنولوجية في التعامل مع هذه البيانات؟,هل تبنى اتخاذ القرارات على ما يستخلص من البيانات ام على الحدس؟,هل يجري تقييم هذه البيانات وإكمالها وتحديثها باستمرار؟,هل ان استخدام تنقيب البيانات يزيد من اداء المصرف؟

أهمية البحث تتمثل في تناول البحث لمتغيرين مهمين تنقيب البيانات DM والاداء يمكن ان يجعل البحث مساهمة متواضعة في هذا المجال خاصة بعد ما لوحظ من نقص كبير جدا في المصادر العربية الخاصة بالمتغير الاول والذي يعد من الموضوعات الحرجة والحديثة جدا "حيث يلاحظ ان اقدم مصدر في هذا الموضوع قد يعود الى عام 1996 كحد أدنى،يساعد البحث المصرف على تجنب مخاطر حقيقية ناتجة عن عدم التسديد والتي تشكل عبئا"على المصرف لكونها قروضا "هالكة،تقديم نماذج تسهل عملية تنقيب البيانات في المصرف حيث تتنبأ بالمقترضين وتصنفهم,حسب مخاطر التسديد لكي يتمكن المصرف من التركيز على المقترضين الموثوق بهم فقط,سعيا منه لزيادة الاداء,يقدم البحث للمصرف قاعدة بيانات يفتقر اليها تخص القروض والمقترضين كما يقدم نظام,متكامل لادارة قواعد البيانات يتمثل بالبرنامج الحاسوبي المستخدم.

يهدف هذا البحث إلي دراسة مشكلة تراكم البيانات ومحاولة إيجاد الحل الأمثل لها،محاولة تحسين مستوى استخدام المعلومات من قبل الادارات من اجل زيادة الاداء,بناء نماذج تنقيب البيانات التي يمكن ان تستخدم في التنبؤ بالمقترضين الموثوق بهم واستبعاد غير الموثوقين منهم لتقليل تكاليفهم و محاولة إرساء أسس سليمة تكون البداية لانطلاق بحوث لاحقة تحاول التطبيق في مجالات اخرى جديدة.

المنهجية المستخدمة: يعتمد البحث على تطبيق برنامج حاسوبي جاهز ومن ثم الوصف والتحليل لنتائج التطبيق من نماذج لاستخلاص المؤشرات الأساسية التي تدعم عملية اتخاذ القرار والرجوع مرة اخرى للتحليل للوقوف على امتدادات نتائج التطبيق.

النتائج التي توصل إليها الباحث هي:

اولا-"الاستنتاجات المتعلقة بالجانب النظري

1. توفر عملية تنقيب البيانات ميزة التعامل مع كميات هائلة من البيانات وكلما كانت كمية البيانات كبيرة كانت اكثر قدرة ودقة في الوصول الى نماذج صحيحة مفيدة.
2. يختلف تنقيب البيانات عن أنظمة دعم القرار كونه قد يكشف عن علاقات ابعد من ان يفكر بها المستخدم.
3. تعتمد نتائج تنقيب البيانات بصورة اساسية على مدى دقة وصحة وشمولية البيانات المستخدمة.
4. تستخدم في تنقيب البيانات تقنيات وادوات عديدة ومتنوعة لكنها تصب في اتجاهين أساسيين هما الاستكشاف والتنبؤ.

لا يمكن تحديد خوارزمية معينة باعتبارها الافضل لبناء نموذج تنقيب البيانات بل يفضل ان تستخدم خوارزميات متعددة ويتم بناء نماذج متعددة الى ان يتم الوصول الى النموذج الافضل.

5. في حالة وجود قاعدة بيانات كبيرة جدا، يكون خيار بناء نماذج عديدة مبنية على عينات افضل من خيار بناء نماذج قليلة مبنية على كل البيانات لكي يكون النموذج اكثر دقة ومتانة.
6. تتطلب عملية قياس الاداء اختيار مؤشرات واضحة سهلة الاستخدام وقليلة الوقت والجهد

ثانيا-"الاستنتاجات المتعلقة بالجانب التطبيقي

1. كلما كان عدد المتغيرات او الاعمدة في قواعد البيانات اكثر كان نموذج التنقيب الناتج اكثر
2. واقعية في التعامل مع مشاكل العمل.
3. سهولة تنقيب البيانات من مصادر بيانات علائقية .SQL Server 2 يوفر برنامج2000
4. ومتعددة الابعاد ومن خلال خوارزميتين هما شجرة القرار والتجميع.
5. يمكن اجراء عملية التنقيب على بيانات تضم متغيرات رقمية وغير رقمية وتظهر في
6. النموذج بالضبط كما صيغتها في قاعدة البيانات مثل مستوى القرض الذي يحمل تقديرات مثل واطيء ومتوسط وعالي.

7. تكون النماذج الناتجة عن عملية التنقيب وخاصة من البرنامج المستخدم سهلة الفهم حيث يستطيع

المستخدم مثل المحلل المالي ان يفسر النموذج ويقارنه مع افكاره الخاصة ,ان سبب ذلك يعود الى ان البرنامج يعمل على اختبار نماذج عديدة ويقدم في النهاية النموذج

8. وفر تطبيق البرنامج عامل السرعة حيث يمكن الوصول الى نموذج التنقيب في غضون دقيقة او اثنتين كحد أقصى مع السماح بتحديث هذه النماذج باستمرار ,ان هذا التحديث والسرعة يسهل على ادارة المصرف اتخاذ القرار الخاص بالتعامل مع طالبي القروض بالرفض او القبول.

9. نتج عن استخدام خوارزمية شجرة القرار نموذج تنبؤي حدد فيه مستوى القرض Level الصفة الأكثر أهمية في التنبؤ بمخاطر القروض ويكون مستوى القرض هنا اما واطئا" او غير واطئ واختارت الخوارزمية بعد ذلك الفرعBranch الصفة الأكثر أهمية الثانية وهو اما ان يكون فرع المصرف في بابل او لا يكون كذلك ويصاحب هذا النموذج وصف واضح ومبسط لكل اجزائه.

10. ينتج عن استخدام خوارزمية التجميع نموذج وصفي حدد عشرة تجمعات مرتبة تنازليا حسب عدد القيود من التجمع الاول وحتى التجمع العاشر، حيث يضم مثلا التجمع السابع اكبر عدد من الحالات السيئة للقروض تصل الى 28 حالة خطر سيئة مقابل 2 حالة خطرة جيدة فقط مع نسب احتمالية كل منها وتنحصر مخاطر القروض السيئة في فروع المصرف في اربيل والسليمانية والمستثمرة في الصناعات الانشائية والميكانيكية ومحطات تعبئة الوقود والنجارة وتكون مستويات القروض متوسط وعالية والمقترضين هم افراد من كلا الجنسين ذكور واناث.

11. ثبوت صحة الفرضية الاساسية حيث ان تطبيق النماذج الناتجة عن خوارزميتي شجرة القرار والتجميع ، يمكن الادارة من رفع نسبة ما مسدد الى ما هو مستحق من قروض ويؤدي ذلك الى تقليل الكلف وزيادة صافي الربح وهو ما يظهر جليا في معيار صافي الربح/ رأس المال حيث حقق المصرف اعلى المعدلات في سنة 2001 ووصل الى699% تقابلها نسبة تسديد عالية نوعا ما مقارنة ببقية السنوات ووصلت الى50.7%

2-1-2 الدراسة الثانية بعنوان بناء **شجرة القرار باستخدام خوارزمية C4.5 لدعم قرارات التسويق المباشر في المصارف لنيل درجة الماجستير بتاريخ 2016/6/13 من جامعة حلب بواسطة** الطالب محمد حسن صديق قسم الاحصاء كلية الاقتصاد جامعة حلب اشراف الدكتور مروان مصطفى ناعسة

مشكلة البحث: تتمثل مشكلة البحث الاساسية في عدم قدرة الاساليب التقليدية على تصنيف العملاء المستهدفين في الحملات التسويقية المباشرة هذا يعني عدم القدرة على التنبؤ باستجابة العملاء الى هذه الحملات وهذا يطرح التساؤلات التالية :

1. هل يمكن استخدام خوارزمية C4.5 في بناء شجرة قرارات تعمل على تصنيف العملاء حول استجابتهم للحملات التسويقية المباشرة ؟
2. ماهو مستوى الإعتماد على هذه الخوارزمية في التنبؤ باستجابة العملاء للحملة التسويقية ؟

اهمية البحث: تاتي اهمية البحث من خلال استخدام تقنية التصنيف كاحدى تقنيات التنقيب في البيانات في تصنيف عملاء المصرف وبالتالي استخلاص قواعد وانماط غير معروفة من كميات كبيرة من البيانات والتي لايمكن اكتشافها بالطرق التقليدية ، والتنبؤ باستجابة العملاء للحملات التسويقية المباشرة مما يزيد في الربحية والحصة السوقية للمصارف من جهة وتخفيض تكاليف الاتصال والوقت من جهة اخرى ,كم تنتج اهمية البحث بتسليط الضوء على اهم طرق التقييم التي يمكن بها الحكم على التقنية المستخدمة ,والاعتماد عليها في الحملات التسويقية المباشرة في المصارف بشكل فعلي وقابل للتطبيق

اهداف البحث: يهدف الى :

1. بناء نموزج شجرة القرار بإستخدام خوارزمية C4.5 ,تعمل على تصنيف العملاء في الحملات التسويقية المباشرة في المصارف
2. تقييم النموزج وفق عدة مقاييس , واختيار قدرته على الوصول للعملاء الذين سوف يستجيبون للحملات التسويقية قبل الاعتماد عليه بشكل فعلي .

منهجية البحث: يعتمد البحث على المنهج التحليلي, حيث يتم تحليل البيانات الخاصة بالعملاء باستخدام الاداة weka لبناء نموزج شجرة قرار للتنبؤ باستجابة العملاء للحملات التسويقية المباشرة في المصارف باستخدام خوارزمية C4.5 ,وسيتم عبر عملية التدريب على جزء من من البيانات تكون فيها حالات الاستجابة اقل بكثير من حالات عدم الاستجابة بغرض تدريب النموزج على البحث عن حالات الاستجابة الايجابية للحملة التسويقية المباشرة ,وتقييم نتائج عملية التدريب ثم بعد ذلك تت عملية اختبار النموذج على جزء آخر من البيانات ,حيث تكون فيها حالات الاستجابة وعدم الاسجابة متساوية بغرض معرفة قدرة النموذج على التفريق بين الحالتين وتقييم نتائج الاختبار للتاكد من دقة النموذج ومدى امكانية الاعتماد عليه في حملات مستقبليه.

النتائج التي توصل اليها الباحث:

1. كفاءة خوارزمية C4.5 على بناء شجرة قرار بغرض دعم القرارات التسويقية المباشرة في المصارف.
2. امكانية تدريب خوارزمية C4.5 على مجموعة من السجلات في ظروف اختلاف كميات الاصنافالمتغير الهدف

3. القدرة على بناء نموزج كامل لبئية تنبؤية عالية ,للتنبؤ باستجابة العملاء للحملات التسويقية.

التوصيات :

1. نوصي المصارف بإستخدام خوارزمية C4.5 في بناء شجرة القرار لتصنيف العملاء حول الاستجابة للحملات التسويقية ,وذلك عبر تكوين قاعدة بيانات كاملة حول العملاء فالظروف المحيطة بشكل مسسبق واستنتاج الانماط التي تحدد سلوكهم تجاه الحمله المستقبلية .
2. استخدام خوارزميات اخرى في بناء نموزج شجرة القرارات لدعم قرارات التسويق المباشر في المصارف وتقييم قدرة النموذج على التطبيق الفعلي.

3-1-2 الدراسة الثالثة ورقة علمية بعنوان استخدام تقنيات تنقيب البيانات لكشف التطفل في شبكات الحاسوب جمهورية السودان , جامعة النيلين، كلية علوم الحاسوب وتقانة المعلومات , الباحثان

أ.إياد محمد مهيوب غالب البريهي – باحث أول

أ.د السماني عبد المطلب أحمد – باحث ثاني

تهدف هذه الورقة لاستخدام تقنيات تنقيب البيانات data mining لكشف حالات الشذوذ في الشبكات تطبيقا على عينتين عشوائيتين من مجموعة بيانات NLS-KDD Data Set ، استخدمت الورقة تقنية التصنيف مصنف شجرة القرار decision tree التي تنفذ خوارزمية C4.5، نتيجة التجربة تعرض أن مصنف C4.5أظهر نتائج فعالة لكشف التطفل في الشبكات الحاسوبية وأظهرت النتائج أنه كلما كانت كمية البيانات كبيرة تكون نسبة الخطأ أقل ودقة التنبؤ عالية , حيث كانت دقة المصنف 63.9 %عندما كانت كمية بيانات الاختبار (946,848 byte) وعندما زادت كمية بيانات إلى (497,700 byte)ارتفعت دقة المصنف إلى (81.5339%) ، بينما كانت كمية بيانات التدريب ثابتة، هذه النتائج تدلل قدرة المصنف العالية للتعلم كلما زاد حجم كمية البيانات.

تمت الاستفادة من هذه الدراسات السابقة في دراستنا الحالية في الآتي

1. الدراسة الاولى: وجود تحليل بيانات والاستفادة منها لاتخاذ قرار مستقبلي كما في حالة القروض هنا هل ثصدق ام لا لجهة ماتكون في حالة الاستيراد لسعات المولدات هل تتم ام لا لزيادة العمليه التسويق
2. الدراسة الثانية: بناء نموزج شجرة القرار بإستخدام خوارزمية C4.5 لتحديد سعات المولدات المستقبلية ذات الشراء العالي وغيرها التي لا يتم شراءها وذلك للاعلان عنها

3. الدراسة الثالثة : زيادة عينات الاختبار كلما كانت كمية البيانات كبيرة تكون نسبة الخطأ أقل ودقة التنبؤ عالية .

2-1-4مقارنة بين الدراسات السابقة والدراسة الحالية:

جدول رقم (1) مقارنة بين الدراسات السابقة والدراسة الحالية

رقم	اسم الدراسة	اسم الباحث	التقنية المستخدمة
1	دور تنقيب البيانات في زيادة أداءالمنظمة(دراسة تحليليه في المصرف الصناعي ببغداد)	إشراف أ .م .د .زكريا مطلك الدوري جامعة بغداد كلية الادارة والاقتصاد	إستخدمت الدراسة تقنية التنقيب في البيانات وخوارزميات شجرة القرار

		الباحثة **داليا عبد الحسين احمد دولة العراق**	و تم ايضا إستخدام خوارزميات التجميع
2	بناء شجرة القرار باستخدام خوارزمية C4.5 لدعم قرارات التسويق المباشر في المصارف	اشراف الدكتور مروان مصطفى ناعسة **جامعة حلب** **الباحث** محمد حسن صديق قسم الاحصاء كلية الاقتصاد جامعة حلب	شجرة القرار باستخدام خوارزمية C4.5
3	ورقة علمية بعنوان استخدام تقنيات تنقيب البيانات لكشف التطفل في شبكات الحاسوب جمهورية السودان , جامعة النيلين، كلية علوم الحاسوب وتقانة المعلومات	أ.إياد محمد مهيوب غالب البريهي – باحث أول أ.د السماني عبد المطلب أحمد – باحث ثاني	إستخدمت الدراسة تقنية التنقيب في البيانات وخوارزميات شجرة القرار
4	الدراسة الحالية: التسويق الذكي بإستخدام التنقيب في البيانات شركة سويتش للتجارة والهندسة(قسم المولدات)	اشراف الدكتورشاذلي صديق محمد احمد الباحثة:سمية عبد الله محمد الامين عبد الله	شجرة القرار باستخدام خوارزمية C4.5

2-2 تنقيب البيانات

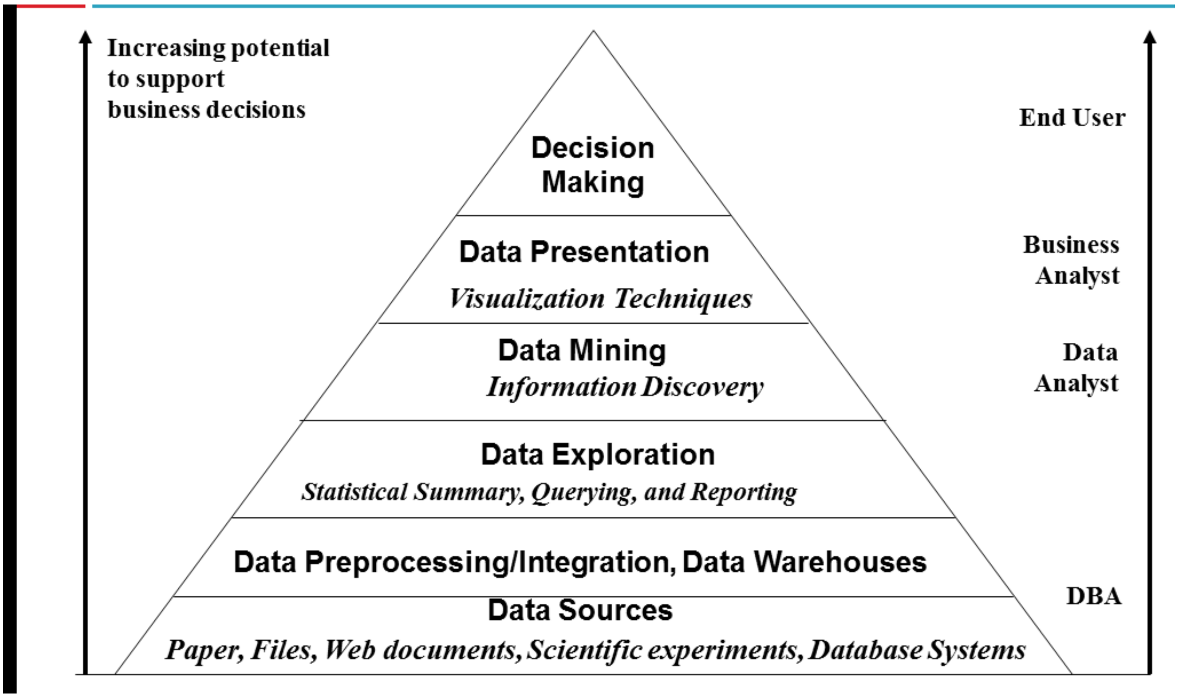
التنقيب في البياناتdata mining هي عملية بحث محوسب ويدوي عن معرفة من البيانات دون فرضيات مسبقة عما يمكن أن تكون هذه المعرفة، كما ويعرف التنقيب في البيانات على أنه عملية تحليل كمية بيانات عادة ما تكون كمية كبيرة لإيجاد علاقة منطقية تلخص البيانات بطريقة جديدة تكون مفهومة ومفيدة لصاحب البيانات، يطلق اسم "نماذج models "على العلاقات والبيانات الملخصة التي يتم الحصول عليها من التنقيب في البيانات, يتعامل تنقيب البيانات عادة مع بيانات يكون قد تم الحصول عليها بغرض غير غرض التنقيب في البيانات (مثلاً قاعدة بيانات التعاملات في للعملاء في شركة ما) مما يعني أن طريقة التنقيب في البيانات لاتؤثر مطلقاً على طريقة تجميع البيانات

ذاتها. هذه هي أحد النواحي التي يختلف فيها التنقيب في البيانات عن الإحصاء، ولهذا يشار إلى عملية التنقيب في البيانات على أنها عملية إحصائية ثانوية .يشير التعريف أيضاً إلى أن كمية البيانات تكون عادة كبيرة، أما في حال كون كمية البيانات صغيرة فيفضل استخدام الطرق الإحصائية العادية في تحليلها.

ان ما يسمى باستخراج البيانات Data Mining هي تقنية تهدف إلى استنتاج المعرفة من كميات هائلة من البيانات، تعتمد على الخوارزميات الرياضية والتي تعتبر أساس التنقيب عن البيانات وهي مستمدة من العديد من العلوم مثل علم الإحصاء والرياضيات والمنطق وعلم التعلم، والذكاء الاصطناعي والنظم الخبيرة، وعلم التعرف على الأنماط ،وعلم الآلة. وغيرها من العلوم والتي تعتبر من العلوم الذكية وغير التقليدية.

وبدأ مفهوم استخراج البيانات DM يظهر الى الوجود في بداية الثمانينات من القرن العشرين وتطور بشكل ملحوظ وانتشر بشكل اكبر في التسعينات من ذلك القرن ويتوقع له الاستمرار بشكل اكثر تطورا وتقدما ، وهو علم يجمع ما بين الإحصاء و تكنولوجيات الإعلام (قواعد البيانات، الذكاء الاصطناعي، التعلم الآلي « machine Learning » ...) ويقدم الحلول الناجحة لتحليل كميات ضخمة من البيانات، وذلك بتحويلها من مجرد معلومات متراكمة وغير مفهومة (بيانات) إلى معلومات قيِّمة يمكن استغلالها و الاستفادة منها بعد ذلك.[4]

وبما ان عملية تجميع البيانات في سجلات بأحجام كبيره مثل (Data warehouse) قد تحتوي على معارف كبيره والتي قد تكون مفيدة لأصحابها من خلال معرفة العديد من الاحصاءات المطلوبة مستقبلا لذلك نحتاج تقنيات مختلفة للاستفادة وهذا ما نريد من التنقيب في البيانات ,فالتقنيات العادية قد لا تكون مفيدة وغير عمليه في مثل هذه الاحجام الكبيرة من البيانات , لذاك تواجدت تقنيات الذكاء الاصطناعي ومنها data mining لتسهل وتحسن عمليه البحث و الاستنباط



Data mining and Business intelligence(1) **شكل رقم**

2-2-1.مخازن البيانات:Data Warehousing

مخازن البيانات أنظمة تستعمل تقنيات جديدة في تخزين كميات كبيرة من المعلومات الغير متشابهة Heterogeneبهدف استعمالها في سرعة اتخاذ القرار لذلك تسمى أنظمة تقريرية Informatique Decisionnelle في مقابل الأنظمة الإنتاجية Informatique de production والتي منها قواعد البيانات ، حتى أن الهيكليات Architecture المستخدمة صممت على استخراج المعرفة للمساعدة في اتخاذ القرار .

برزت الحاجة الملحة إلى مخازن البيانات في الشركات الكبرى حيث أن كل قسم من هذه الشركات يدير قواعد بيانات مستقلة خاصة به (تسويق، مالي، ...) والتي تتضمن كثيراً من البيانات المشتركة وتتطلب حاجة الشركة إيجاد كامل البيانات المتعلقة بموضوع محدد كالزبائن مثلاً من خلال بحث في قاعدة واحدة بدلاً من البحث في القواعد المختلفة .

لذا برزت الحاجة إلى تجميع البيانات في قاعدة واحدة تسمى مخزن البيانات مع الاحتفاظ بالقواعد الخاصة على أن يصار إلى تغذية المخازن دورياً في حال حصل تعديل أو تبديل في القواعد وتصنف البيانات بحسب الموضوع وذلك باستخدام برامج خاصة تسمى Middleware تستعمل بيانات دمج خاصة في كل موضوع تسمى Schema d integrationتحدد فيها شكل وكيفية نقل البيانات .

مستودعات البيانات عبارة عن مكان للبيانات الثانوية التي نظمت من التطبيقات الاخرى او من

مصادر او نظم خارجية يتم خزن استفسارات قواعد البيانات المثلى وادوات التقارير بسبب قدرتها على تحليل البيانات غالبا من قواعد بيانات متباينة وطرق مفيدة . انها وسيلة للمدراء وصناع القرار لاستخراج المعلومات بسرعة وبسهولة للاجابة على الاسئلة حول اعمالهم وبعبارة اخرى مستودعات البيانات هي قواعد بيانات متكاملة للقراءة فقط صممت لعقد المقارنات والاجابة على سؤال ماهي ؟وبالاختلاف مع قواعد البيانات التي تمسك بالتعاملات وتحتفظ بها لبيانات مشابهة لاخر معاملة، مستودعات البيانات هي تحليل موضوع – توجيه وهيكل او تركيب لمجموع المعاملات كلقطة في الوقت المناسب

معالجة معلومات الدعم بواسطة تزويد برنامج صلب موحد ، بيانات تاريخية للتحليل ، تصميم خاص للاستفسارات والتقارير فضلا عن معالجة المعاملات او الصفقات .

أ. **خصائص مستودع البيانات data warehouse**

1. هي عبارة عن قواعد البيانات
2. تسمح بتكامل نظام التطبيقات المتنوع
3. تدعم معالجة المعلومات (المعالجة التحليلية) .

ب. **صناعة مخازن البيانات :**

أما كيفية صناعة مخازن البيانات فتتم عبر :

– توحيد البيانات الواردة من مصادر متنوعة وغير متشابهة (قواعد البيانات)
–توزيع البيانات بحسب المواضيع والاهتمام (أقسام الشركة مثلاً)
–الاهتمام بالبعد الزمني(التاريخ)
–حفظ كامل البيانات من الأقدم حتى الأحدث Historique

ج. **عرض نتائج البحث في مخازن البيانات :**

أما في عرض نتائج البحث فتستخدم طريقتين لتحليل البيانات ، فالطريقة الحسابية تستعمل تحليل البيانات من أجل إظهار النتيجة بطريقة حسابية إحصائية (الجمع، المعدل،...) أو استخراج البيانات Data Mining وإظهار البيانات بشكل Diagram تحوي كل واحدة منها معلومات في مجال محدد (شؤون الموظفين، التسويق،..) تستعمل في تحديد نطاق البحث واستعمال أنظمة التحليل على هذه المخازن لاتخاذ القرار في نطاق محدود

2-2-2اهداف التنقيب

هناك ثلاثة أهداف للتنقيب في البيانات:

1) من أجل تعليل بعض الظواهر المرئية.

2) من أجل التثبت من نظرية ما. مثال: التثبت من النظرية التي تقول بأن الأسر الكبيرة تهتم بالضمان الصحي أكثر من الأسر الصغيرة عددا.

3) من أجل تحليل البيانات للحصول على علاقات جديدة وغير متوقعة. مثال: كيف سيكون الانفاق العام إن كان ملازما لعمليات خداع واسعة من قبل البطاقات الائتمانية.

2-2-3وسائل التنقيب في البيانات

هناك عدة وسائل مختلفة من أجل التنقيب في البيانات اختيار الوسيلة المناسبة يعتمد على طبيعة البيانات تحت الدراسة وعلى حجمها يمكن اجراء عملية التنقيب في البيانات بالمقارنة مع سوق البيانات ومخزن البيانات فوسائل التنقيب في البيانات تُستعمل وبنجاح في الكثير من التطبيقات الحقيقة حول العالم.

جدول رقم (2) امثلة لبعض وسائل التنقيب (يتبعها مثال لكل تطبيق) هي:

الوسيلة	يتبعها وظيفة كل وسيلة
الاستدلال المبني على حالات سابقة Case-Based reasoning	التفكير واستخلاص النتائج والقوانين من أمثلة حية و قضايا تم حلها سابقا
الكشف عن قانون Rule Discovery	البحث عن منوال معين أو علاقة معينة في جزئية كبيرة من البيانات
معالجة الإشارات Signal Processing	ايجاد الظواهر المتشابه مع بعضها البعض
شبكات عصبونية Neural Nets:	تطوير نماذج قابلة لتنبؤ النتائج. هذه النماذج تم تطويرها بناء على أسس تم استنباطها من عقل الإنسان
منحنيات غير ثابتة Fractals	تصغير البيانات الكبيرة من دون ضياع المعلومات
كتابة تقرير مختصر عن فئة معينة Profiling Populations	تطوير وإنشاء تقارير موجزة عن الزبائن المهمين وعن بطاقات الائتمان
تحليل النزعة التجارية Analysis of Business Trend	ايجاد الأسواق ذات قدرات النمو القوية أو الضعيفة
التسويق لفئة معينة Target Marketing	ايجاد الزبائن من أجل منح التخفيضات لهم لسبب معين
تحليل الاستعمال Usage Analysis	ايجاد منوال معين لاستعمال الخدمات والسلع
فعالية الحملة Campaign Effectiveness	مقارنة استراتيجيات الحملات مع بعضها البعض من أجل ايجاد أكثرها فعالية وتأثيرا
جاذبية السلعة	ايجاد السلع التي تباع مع بعضها البعض

.2-2-4طرق ونسق التنقيب في البيانات

1-طريقة التنبؤ (prediction method) :- استخدام البيانات المتوفرة وتطبيق عليها تقنيات معينه لتحقيق واعطاء قيم مستقبليه ناجحة ...

2-طريقة الوصف (Description method) :- عملية وصف للبيانات المتاحة ومعرفة تصنيفاتها حسب تواجدها والعلاقات بينها

من خلال المحاكاة للروابط الطبيعية (human interpretable) بمعنى اخر نأخذ الروابط من خلال التفاعل الطبيعي لكي نشرح هذه البيانات.

5-2-2الفرق بين عملية الاستعلام العاديه على قواعد البيانات العاديه والاستعلام في البيانات المنقبه

المثال التالي والذي يوضح الفرق في الاستعلام بين كلا النوعين ..

Data base Query vs. Data mining Query

الاستعلام عن معلومات في سوق تجاري ....

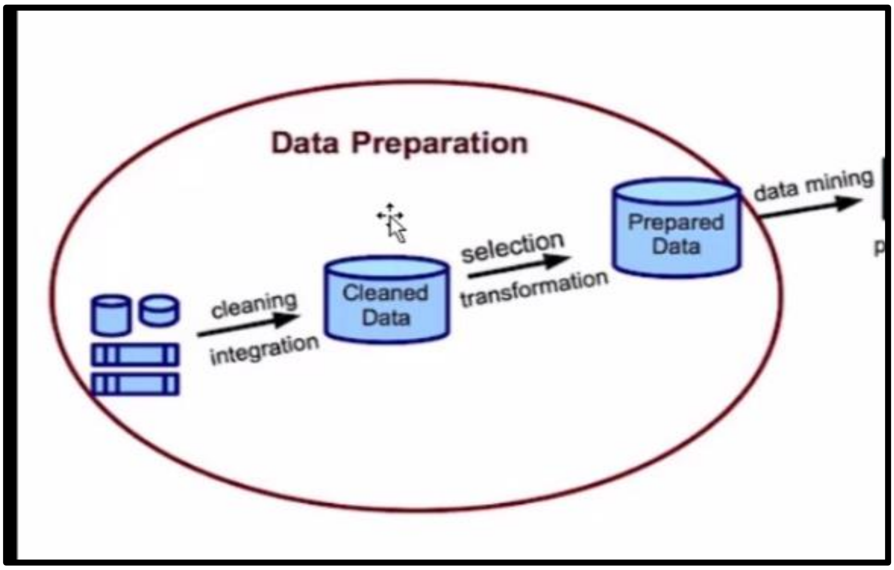
في قاعدة البيانات العادية (Data base)

1. الاستعلام عن المعلومات الكاملة للزبائن الذين ينتهي اسمهم الثاني بخالد ..
2. الاستعلام عن الزبائن الذين اشتروا بمبلغ اكثر من 1000 دينار في اخر الشهر.
3. الاستعلام عن اسماء الزبائن الذين اشتروا الحليب .

اما في حالة الاستعلام في (Data mining)

1. الاستعلام عن كل البطاقات الائتمانية والتي عليها مشاكل .. وهذا يسمى (classification) التصنيف .
2. الاستعلام عن الزبائن الذين لديهم عادات شرائية مماثله .. وهذا يسمى (clustering) التجمع او العناقيد .
3. الاستعلام عن السلع التي يتم شراءها بشكل تزامني مع الحليب , المقصود في كل عملية شراء يقوم بها الزبائن ..وهذه تسمى (association rules) اكتشاف قواعد وعلاقات الارتباط .

2-2-6.تهيئة البيانات (Data preparation)



شكل رقم (2) تهيئة البياناتdata preparation

2-2-7-1تنظيف البيانات(data cleaning).

هنا يتم التخلص من بعض البيانات التي تحتوي على عيوب اما تكون مطبعيه , او بيانات قديمة لا تفيد في الوقت الحالي وغير ذلك .

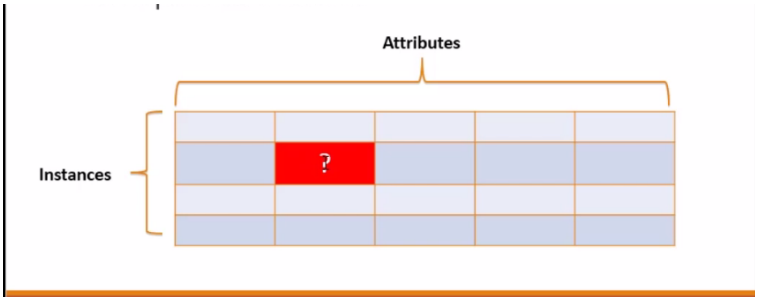
بعض العوامل المتسببة في تقليل جودة البيانات

1. **Noise** يشير مصطلح الضوضاء أو Noise إلى التغييرات التي تطرأ على قيم البيانات الأصلية كالتشويه الملحوظ في صوت محدثك على الهاتف،إذا ما زادت نسبة الضوضاء في البيانات دل هذا علي فسادها و عدم صلاحيتها للاستخدام أما إذا كانت النسبة في الحدود المعقولة فينبغي عندئذ استخدام تقنيات التنقيب المختلفة.
2. **Outliers** و هي بيانات ذات خصائص مختلفة عن باقي البيانات في مجموعتها، و تختلف Outliersعن Noise حيث يُتطلب منا اكتشاف المتغيرات Outliers – بعكس الضوضاء Noise –التي يُتطلب منا توضيحها و تنقيتها.
3. **Data duplication** قد تكون البيانات مكررة.
4. **Timeliness** بعض البيانات تفقد بمرور الوقت أهميتها.

جدول رقم (3) Data Cleaning

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes		125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95k	Yes
6	No	Married	60K	No
7	Yes		220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

2-2-7-2 البيانات المفقودة : عمليه تنقيب البيانات من اهم متطلباتها ان تكوت البيانات كامله لا تحتوي على قيم مفقودة , وهنالك طرق طبعا لتصحيح هذه البيانات , مثل الوسط الحسابي وغيرها من العمليات (Missing value) .



شكل رقم (3) تمثيل القيمة المفقودة

2-2-7-3 اشتقاق البيانات :- في بعض الاحيان يستوجب اشتقاق بعض الأعمدة فتساعدنا في الحصول على معلومات مفيدة (Data derivation) .

2-2-7-4 دمج البيانات :- في بعض الاحيان يتم دمج بعض الأعمدة للحصول على نتائج افضل او الاختصار في البيانات (Merging Data) .

2-2-8تطبيقات التنقيب في البيانات

تطبيقات التنقيب في البيانات بدأت تنمو بصورة كبيرة للأسباب التالية:

1) كمية البيانات الموجودة في مخزن البيانات وسوق البيانات تنمو بصورة أسية (exponential).

ومن أجل ذلك، فإن المستخدم يحتاج إلى أدوات متطورة من مثل التنقيب في البيانات من اجل استخلاص الفائدة والمعرفة من هذه البيانات.

2) الكثير من أدوات التنقيب عن البيانات بدأت تظهر مؤخرا، وكل أداة أفضل من الأخرى.

3) المنافسة الشديدة الموجودة في السوق تدفع الشركات إلى الاستفادة القصوى من البيانات التي بيدها. عمليات التنقيب في البيانات تفعل ذلك تماما.

2-2-9أمثلة عن أهمية تطبيق تقنيات التنقيب في البيانات

يمكن بيان أهمية تطبيق تقنيات تحري البيانات في مثال بسيط يعرض وضع شركة توزيع، تتلقى طلبات شراء من المواطنين، تُدخل هذه الطلبات في نظام معلوماتي قبل إرسال المواد والبضائع التي تتضمنها هذه الطلبات إلى الزبائن. بعد العمل مدة طويلة بهذه الطريقة، سوف يتوافر لدى هذه الشركة قاعدة معطيات مهمة تحتوي على المعلومات التي تتضمنها طلبات الزبائن في تلك المرحلة، وسوف تكون هذه القاعدة مصدراً مهماً للمعلومات، تستفيد منه الشركة في اتخاذ القرارات المناسبة لعملها، ووضع خططها المستقبلية وتحديد مجموعات الزبائن التي تتعامل معها، ومن ثمّ تصنيفها وفقاً للمناطق السكنية، أو الفئات العمرية، أو المنتجات المطلوبة، أغيرها من المعايير البسيطة أو المركبة. كما تستطيع الشركة دراسة وسائل التسويق والدعاية التي تستخدمها، ومردود هذه الوسائل. فقد تكتشف الشركة أن عدد زبائنها الذين تعرفوا على منتجاتها من الرسائل البريدية التي أرسلتها إلى سكان منطقة سكنية معينة لم يتجاوز نسبة 4٪، وأن الربح الذي حققته من بيع منتجاتها لهذه الشريحة من الزبائن لا يغطي ما أنفقته في إرسال هذه الرسائل، ومن ثمّ لابد من إيجاد وسائل ترويج أكثر ملاءمة مثل الإعلان عن تخفيضات في أسعار بعض المنتجات أو توفير طرق تسديد تناسب مستوى دخلهم. كما يمكن تطبيق تقنيات أكثر تعقيداً مثل الشبكات العصبونية لمحاكاة عملية إرسال الإعلانات بالبريد وتجاوب الزبائن مع هذه الرسائل. تتعلم الشركة من المعلومات المسجلة حول ما جرى في الماضي ما يرشدها بعد ذلك إلى اختيار العناوين التي يمكن أن تتجاوب مع هذه الطريقة في الإعلان.

كذلك يمكن الانتقال إلى مستوى أكثر تعقيداً من دراسة الزبائن والعلاقة التي يمكن أن تكون بين مجموعات الزبائن ومجموعات المنتجات التي تسوقها الشركة؛ فالزبائن الذين يشترون اليوم ملابس لأطفالهم الرضع، قد يشترون بعد عشر سنوات الألعاب الإلكترونية لهؤلاء الأطفال.

يتبين من هذا المثال أهمية معرفة هذه الأنواع من القواعد وتطبيقها لإيجاد فرص تجارية أفضل، ومعرفة مدى فعالية طرق التسويق التي تطبقها وتدقيقها، ومقارنة هذه الوسائل وتحديد ما يناسب شريحة معينة من الزبائن.

لا يقتصر التنقيب في البيانات على التطبيقات من النوع السابق، وإنما يشمل طيفاً واسعاً من النشاطات، فمثل هذا التقنيات تفيد، إذا ما طُبقت على قواعد البيانات التي تحوي معطيات عن المشافي و المرضى الذين يدخلون، والأدوية التي يتعاطونها وطرق العلاج المستخدمة، ومدى استجابة الأمراض لها، والتوزع السكاني للمرضى، وأحوال الطقس في هذه المناطق السكنية، والأعمال التي يمارسها سكانها، ومستواهم التعليمي في معرفة الأمراض التي تصيب فئة معينة منهم، واكتشاف الأوبئة التي تصيب الناس في أوقات معينة واتخاذ الاحتياطات اللازمة لها، وغير ذلك من المعلومات؛ تفيد في اتخاذ قرارات تتعلق بتحسين مستوى الخدمات الطبية التي تقدمها هذه المشافي.[5]

2-2-10كيفية دراسة البيانات

وهذه الطريقة بالأساس تهمنا لكي نحدد طريقة تعلم هذه البيانات فأما ان تكون :-

(supervised learning) التعلم الاشرافي وهنا يكون بوجود هدف منشود والخوارزمية تحاول الوصول به من خلال مجموعه من المتغيرات او المعطيات .او بمعنى اخر ان الخوارزمية المستخدمة تقوم بعمل الربط بين الهدف المنشود والمحدد سابقا مع ما يرتبط مع من الحلول والاقتراحات الموجودة في مسار الحل (التعلم) مقارنة المتوقعة مع الهدف المحدد اصلا .(التعلم من خلال امثله)

مثال :- classification

*(Un-Supervised learning) وهنا تقوم بتجميع البيانات على شكل مجموعات تحتوي على بيانات متشابهة الخصائص

والك لتمييز بعض الاستثناءات بينها .(التعلم من غير معرفه مسبقه عن الحل الصحيح المطلوب)

مثال:- clustering .

كما قلنا سابقا اول خطوه في تنقيب عن البيانات كانت تهيئتها لننتقل لباقي الخطوات

الخطوه الثانيه :- ندرس الحاله ونتعرف عليها وطبعا نحدد هل تمثل تعلم اشرافي او غير اشرافي كما وضحنا في الدرس السابق وهنالك نقاط مهمه يتم الاهتمام بها وسالخصها بنقطتين :-

1:- تحديد او التعريف بالدراسه التي نريد ان نقوم بها ينطوي على تحديد مجال قاعدة البيانات المستخدمه !!!!!!

مثال :- اول مجموعة بيانات لبناء النموذج وقاعدة بيانات اخرى للتحقق من صحة هذا النموذج ومجموعه اخر لايجاد التكهنات او التوقعات لهذا النموذج .. (التدريب , والتحقق ومجموعة الاختباراو الفحص) .

تحديد حجم العينة , وذالك لانه ليس بالضروره اجراء عملية التنقيب على كل البيانات الموجوده , فيمكن اختيار مجموعة من الصفوف من خلال عينات عشوائيه .

الخطوه الثالثه :- قراءة البيانات وبناء النماذج

والنموذج يلخص حجم كبير من البيانات من خلال عدة مؤشرات كالاتي :-

من خلال الترددات او التكرارت :- فهو غالبا يظهر حدوث قيمه معينه ويبين كم نسبة التاكد من القيمه المحدثه .

من خلال الوزن او التاثير :- يظهر كيف ان بعض المدخلات تشير الى حدوث المخرجات , اي بمعنى اخر يوضح ان بعض الاعمده مثلا لها اوزان عاليه .. يبين كيف المدخلات المشار اليها تعطي المخرج المطلوب.

من خلال الارتباطات :- بعض المدخلات لها اوزان عاليه عندما تكون مع بعض افضل من اوزانها عندما تكون منفرده .

التمايز او المفاضله :- تبين مدى اهمية اوزان ومعاير بعض المدخلات في المخرجات بالنسبه لغيرها من المدخلات .

من خلال تحديد الاختلاف في الاوزان والمعايير بينها ..

الخطوه الرابعه :- فهم النموذج التالي والذي يمثل مجموعة بيانات اخذت من دراسة لارتفاع ضغط الدم ..

حيث بعض النماذج تمثل على اشكال ومنهجيات مختلفة كالصورة السابقة والتي كانت Decision tree

حيث تم فصل البيانات على اساس صفة العمر .

او مثلا نموذج الشبكة العصبية والذي انتقد على اساس انه الصندوق الاسود لتوسعه في استخدام مبدأ التنبؤ . لنفرض ان احدى البيانات نريد ان نجري لها تنقيب , فمراحل فهم النموذج الذي أنشئ من هذه القاعدة من البيانات تشمل الاتي :-

1-ملخص النموذج: - يظهر الاوزان والترددات والارتباطات المهمة لكي توضح الهدف .

2-توزيع البيانات: -فمن المهم ان تكون البيانات كبيره كفاية لتمثيل العينة .

3- الفحص والمقارنة : وتنطوي على استخدام التنبؤات التي يحققها النموذج المستخدم ومقارنة النتائج الجديدة بالنتائج المعروفة .

4- التحقق من فشل او نجاح التنبؤ :- معرفة لماذا النموذج فشل او نجح في التنبؤ .

5- ومن ثم تحديد الصفوف التي تعتبر الافضل والاكثر حضا لتحقيق النتائج المرجوة .

الخطوة الخامسة : التنبؤ (Prediction) هي عملية التنبؤ بالنتائج بناءً على البيانات الموجودة .[6]

تنقسم البيانات عادةً إلي ثلاثة أقسام:

1. بيانات في صورتها الخام دون أي فحص أو تحليل و تُسمى.Data
2. بيانات تم تحليلها و استخلاص بعض المعلومات البسيطة منها و تُسمى.Information
3. معلومات تم استخلاصها بطرق أكثر تعقيداً و إذا ما تم إضافتها إلي خبرة المُحلّل يصبح بحوزتنا ما يُسمي.Knowledge

2-2-11عملية التنقيب: Mining –

- **التحليل المبدئي و التحويل-: Preprocessing & transformation –**

يتم في تلك المرحلة تحويل البيانات الخام المختلفة في شكلها و صياغتها إلى شكل عام موحد لمرحلة التحليل اللاحقة، تُعد هذه المرحلة هي الأكثر استهلاكاً للوقت و الجهد.

- **التصفية و التقييم-: Evaluation & filteration –**

تتضمن هذه المرحلة التأكيد علي صلاحية و فائدة النتائج المستخلصة كي يتم دمجها في عملية صنع القرار.

تواجه عملية التنقيب عدة تحديات نوجزها فيما يلي-:

- **القدرة علي الاستيعاب و التوسع-: Scalability –**

و تقيس مدي قدرة الخوارزميات الموجودة حالياً على معالجة القدر الضخم و الهائل من البيانات.

- **البعدية الكبيرة-: High dimensionality –**

تحتوي البيانات الآن علي آلاف الخصائص المختلفة بالتالي لن تستطيع الطرق التقليدية للتحليل معالجتها بشكل سليم.

- **البيانات المعقدة و المتغيرة الخواص-: Heterogenous and complex data –**

تتعامل الطرق التقليدية للتحليل عادةً مع بيانات ذات سمات موحدة ، علي عكس بيانات الحمض النووي علي سبيل المثال.

- **ملكية البيانات-: Data ownership –**

احياناً نحتاج إلى بيانات غير مُخزنة في مكان ثابت و لا تنتمي لمؤسسة بمفردها مما يتطلب منا تطوير تقنيات لتوزيع تلك البيانات بأمان.

- **التحليل غير التقليدي:- Non- traditional analysis –**

العديد من طرق التحليل تتطلب إبداع و تقييم الالاف من الفرضيات و بالتالي لابد من تحديث تقنيات التنقيب في البيانات من أجل تطوير تلك الفرضيات آلياً في المستقبل.

كي يؤدي التنقيب في البيانات مهمته بنجاح، تم تزويده بمجموعة خوارزميات كل منها لها وظيفة و دور محدد تقوم به، يُعتبر اختيار الخوارزمية الصحيحة عمل ليس سهلاً على الإطلاق فكل خوارزمية تُظهر نتيجة مختلفة عما إذا كنت قد استخدمت خوارزمية أخرى

وتتألف تقنية استخراج البيانات من عدد من الاساليب منها قواعد الارتباط Association Rulesوالعنقدة clustering والتصنيف .. Classification الخ . وباستخدام هذه التقنية يمكن استنتاج المعرفة المختفية خلف العديد من البيانات والمعلومات والتنبوء والقدرة على اتخاذ القرارت

1.Classification

يُستخدم في تحديد مدي مطابقة الفرضيات الجديدة لأي من الفئات المحددة سلفاً.

بفرض اننا نريد كشف جرائم الاحتيال بالنسبة لبطاقات الائتمان، في البدء نستخدم العمليات السابقة التي أُجريت للبطاقة مثل : المنتجات التي يشتريها صاحبها عادةً، هل يدفع في الأوقات المحددة أم لا ؟ ، ثم نصنف تلك العمليات باعتبارها صحيحة أم لا، بعدها نحدد نمط معين تجري فيه مثل هذه التعاملات .. (كأن يدفع صاحبها اشتراك النادي ثم يدخل سينما ما بعدها) .. نستخدم هذا النمط فيما بعد لكشف الاحتيال المصرفي.

2.Regression

تقيس مدي الارتباط بين المتغيرات المختلفة، فهي تتنبأ بقيمة متغير معين بناءً على قيم متغير أخر.

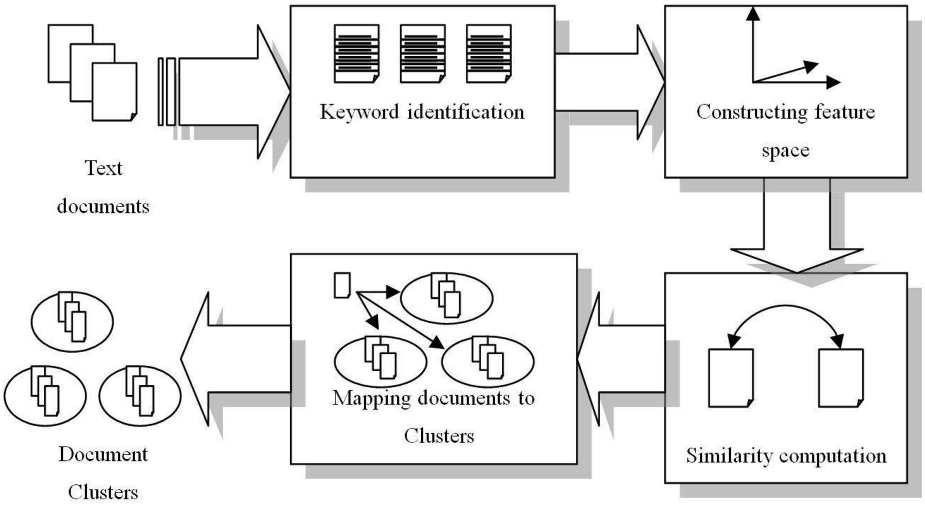
علي سبيل المثال : توقع عائد بيع منتج جديد بناءً علي قيمة إعلاناته.

3.Clustering

تسعي تلك الطريقة الى تصنيف المعلومات الى كتل متشابهة في خصائصها بالتالي كل كتلة تكون متشابهة في خصائصها عن الكتلة الأخرى.

مثال علي ذلك هو Document clustering ، الهدف هنا هو تصنيف الوثائق الي مجموعات متشابهة بناءً علي أهمية المصطلحات الواردة بها بالتالي يتم تقييم المصطلحات الواردة

باستمرار في كل وثيقة مع إنشاء وحدة قياس تماثلية Similarity measure – تحدد عدد الكلمات المشتركة في تلك الوثائق.



شكل رقم (4) توضيح عملية تنقيب نصوص باستخدام التجميع Document clustering

Sequential pattern discovery.4

وهواكتشاف نمط متكرر بترتيب معين كشراء ادوية للبرد يتبعها اغذية معينة تساعد علي تخفيف المرض.

Anomaly detection.5

وهو كشف الانحرافات الملحوظة عن السلوكيات المعتادة ، مثال علي ذلك هو التطفل الشبكي Network intrusion – فقد نلحظ احياناً زيادة مفاجئة و عالية علي اتصال شبكي ما مما قد يدل علي وجود هجمات الحرمان من الخدمات Denial of service attack – أو اختصاراً . DOS

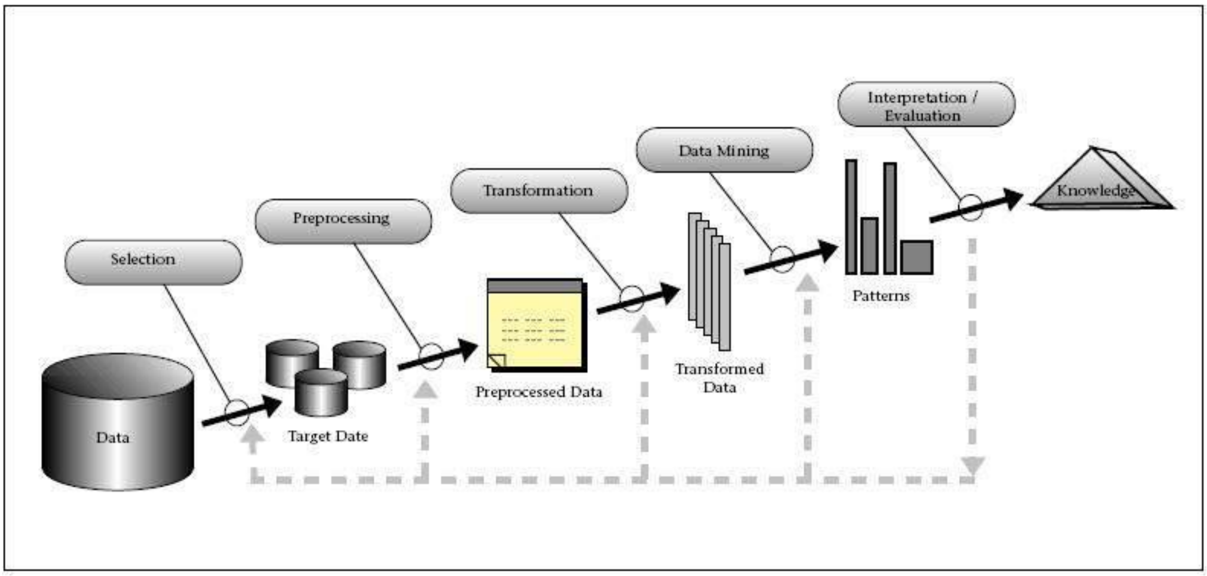
تركز عملية التنقيب عامةً علي اكتشاف و حل المشاكل المتعلقة بجودة البيانات و كذا استخدام الخوارزمية الافضل في التعامل مع بيانات ذات جودة سيئة ، فمن غير المنطقي افتراض وجود بيانات ذات جودة عالية حيث توجد عوامل كثيرة قد تحول دون ذلك كالأخطاء البشرية. كل ما يظهر لك الآن من إعلانات و عروض سلعية بل حتى اقتراحات الصداقة على مواقع التواصل الاجتماعي تتم بحرفية و عن طريق جمع معلوماتك المتناثرة في الفضاء السيبراني يمكن تقريباً معرفة الكثير و الكثير عنك و أحياناً تُباع تلك المعلومات للشركات بغرض دراستها و اعتصار معلومات دقيقة منها تجبرك

في النهاية على الخضوع لإغراءات الشراء و التبضع بعد معرفة خباياك و أمنياتك، نحن في عالم أصبحت الخصوصية فيه وهم يسكن عقول المؤمنين به فقط.[7]

12-2-2 مراحل اكتشاف المعرفة :

Knowledge Discovery in Database (KDD) ليس بالعملية السهلة والتي قد يعتقد البعض أنها تتوقف عند تجميع البيانات وإدارتها، بل نراها تمتد إلى التحليل والتوقع والتنبؤ بما سيحدث مستقبلاً.

التنقيب في البيانات يشكل جزءا من اكتشاف المعرفة knowledge discovery، وهذه العملية هي الأكثر شمولا.



شكل رقم (5) رسم يوضح مراحل اكتشاف المعرفة

تتضمن عملية اكتشاف المعرفة الخطوات التالية:

١- اكتشاف البيانات Data discovery:

وهي مرحلة جمع البيانات وتشمل كشف وتحديد وتوصيف البيانات المتاحة.

٢- تصفية البيانات وتنقيتها Data cleaning:

ويتم في هذه المرحلة إزالة البيانات المزعجة Noise التي لا أهمية لها، كما يتم حذف البيانات المتضاربة والبيانات الغير متناسقة.

٣- تكامل البيانات Data integration:

يتم في هذه المرحلة تجميع البيانات المتشابهة وذات الصلة من مصادر البيانات المتعددة ودمجها معا

٤- اختيار البيانات Data selection:

في هذه المرحلة، يتم تحديد واسترجاع البيانات الملائمة من مجموعة البيانات.

٥- تحويل البيانات Data transformation:

في هذه المرحلة يتم تحويل البيانات إلى نماذج مخصصة ملائمة لإجراءات البحث والاسترجاع بواسطة خلاصة الإنجاز أو عمليات التجميع.

٦- التنقيب عن البيانات Data mining:

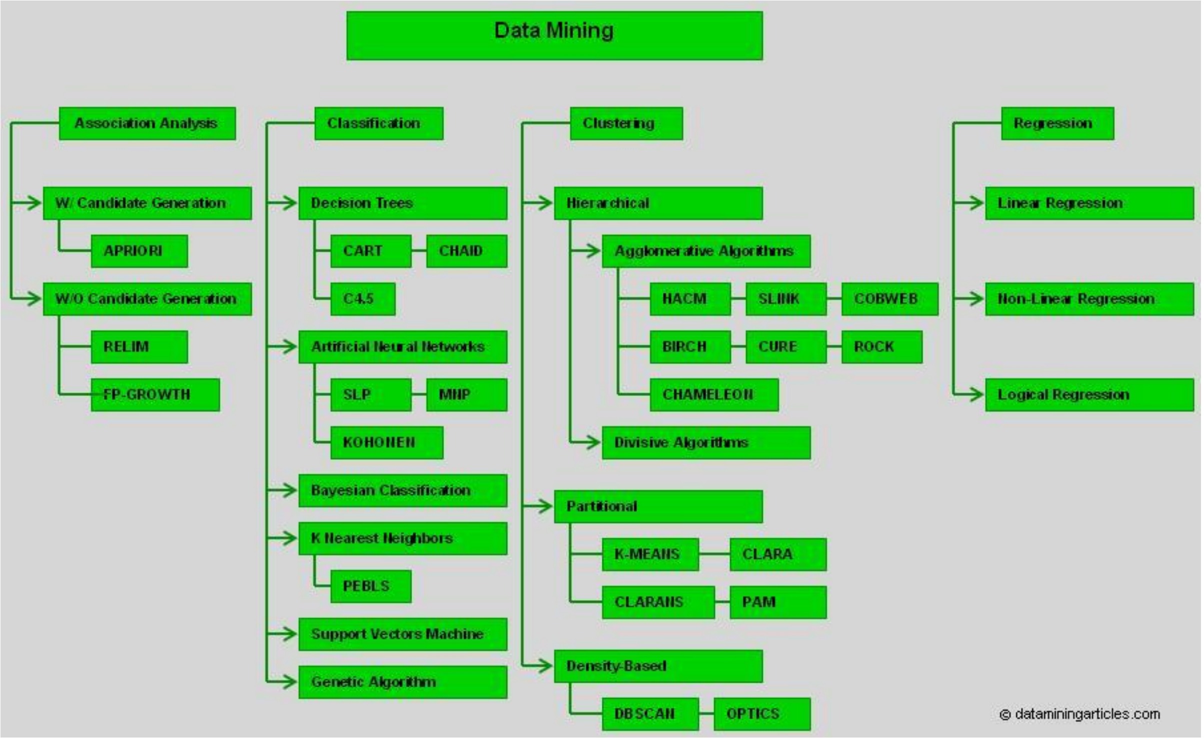
أي استخدام طرق ذكية تطبق لاستخلاص أنماط البيانات استخراج نماذج مفيدة قدر الإمكان.

٧- تقييم النمط Pattern evaluation:

يتم في هذه المرحلة تحديد الأنماط المهمة حقا والتي تمثل قاعدة المعرفة لاستخدام بعض المقاييس المهمة.

٨- تمثيل المعرفة وتقديمها Knowledge presentation:

وهي المرحلة الأخيرة من مراحل اكتشاف المعرفة في قواعد البيانات وهي المرحلة التي يراها المستفيد، هذه المرحلة الأساسية تستخدم الأسلوب المرئي لمساعدة المستفيد في فهم و وتفسير نتائج استخراج البيانات. وبالتالي التنقيب في البيانات هو خطوة أساسية لتطبيق أساليب ذكية بهدف الكشف عن أنماط البيانات المثيرة للاهتمام والمخبأة في مجموعات البيانات الكبيرة. ومع ذلك، في بعض المنظمات نجد أن مصطلح التنقيب عن البيانات "data mining" أصبح أكثر شعبية للإشارة إلى العملية التي يتم فيها اكتشاف المعرفة knowledge discovery برمتها. وهناك جانب مهم جدا، يجب النظر إليه بعين الاعتبار و هو أن هناك أنماط جديدة قد تبرز ،عادة ما تكون غير معروفة من قبل. وبالتالي يجب أن تكون أدوات التنقيب عن البيانات قادرة على البحث عن أنواع مختلفة من الأنماط، بأشكال متوازية لزيادة كفاءة التنقيب عن البيانات. كما يجب أيضا أن يتم الكشف عن الأنماط في الأجزاء الصغيرة والفرعية، والتي تعرف بالحبيبات granularities ، مما يعني البحث في مستويات مختلفة من التجريد أو التفصيل. حلول التنقيب عن البيانات الجيدة هي التي تشير أيضا إلى قدر من الثقة أو اليقين المرتبطة مع نمط اكتشافها، لأن بعض الأنماط قد لا تصلح لكافة البيانات في مجموعة البيانات التي تم تحليلها.[8]



شكل رقم (6) رسم يوضح اساليب تعدين البيانات

الشكل السابق يوضح تقنية استخراج البيانات من عدد من الاساليب منها قواعد الارتباط Association Rules و العنقدة clustering والتصنيف .. Classification الخ وكل اسلوب له عدة تقنيات وباستخدام هذه التقنية يمكن استنتاج المعرفة المختفية خلف العديد من البيانات والمعلومات والتنبوء والقدرة على اتخاذ القرارت

2-3-1أهم طرق تنقيب البيانات:

هناك وسائل مختلفة للتنقيب في قواعد البيانات اختيار الوسيلة المناسبة يعتمد على طبيعة البيانات تحت الدراسة وعلى حجمها يمكن إجراء عملية التنقيب:

2-3-2التصنيف Classification :

التصنيف هو شكل من أشكال تحليل البيانات التي تستخرج النماذج التي تصف طبقات البيانات الهامة , تعرف هذه النماذج بالمصنفات , تتوقع Class Labels.

على سبيل المثال يمكننا بناء نموذج تصنيف لتصنيف طلبات الحصول على القروض المصرفية أما أمنة (Safe) أو خطرة (Risky) , هذا التحليل يمكن أن يوفر لنا فهما أفضل للبيانات ككل.

يستخدم التصنيف بشكل واسع في حل كثير من المشكلات من خلال تحليل مجموعة من البيانات ووضعها في شكل أصناف أو أقسام يمكن استخدامها فيما بعد لتصنيف البيانات المستقبلية.

التصنيف له تطبيقات عديدة بما في ذلك (الكشف عن الغش Fraud Detection – التسويق Marketing – التصنيع Industry والتشخيص الطبي) [9].

2-3-3خطوات التصنيف :

التصنيف يتم في خطوتين

الخطوة الأولى : خطوة التعلم ويتم بناء نموذج التصنيف , حيث خوارزمية التصنيف تبني المصنف Classifier من خلال التعلم من مجموعة تدريبية (Training Set) تتكون من Database & Class Labels Tuples المرتبطة بها , مجموعة التدريب (Training Set) تسمى (Training Tuples) ويتم أخذها بشكل عشوائي من قاعدة البيانات قيد التحليل أو المراد تصنيفها.

الخطوة الثانية : تحديد ما إذا كانت دقة النموذج مقبولة وإذا كان الأمر كذلك , سوف يتم استخدام النموذج لتصنيف البيانات الجديدة بمعنى استخدام النموذج للتنبؤ بالـ Class Labels لبيانات معينة.

2-3-4طريقة تعلم البيانات :

التعلم الإشرافي Supervised Learning : وهذا يكون بوجود هدف منشود والخوارزمية تحاول الوصول به من خلال مجموعة من المتغيرات أو المعطيات أو أن الخوارزمية تقوم بعمل الربط بين الهدف المنشود والمحدد مسبقا مع ما يرتبط من الحلول والاقتراحات الموجودة في مسار الحل (تعلم مقارنة المتوقعة مع الهدف المحدد أصلا، مثال Classification.

يعني التعلم المشرف : إذا كان الـ Class Labels لكل صفوف التدريب Training Tupels متوفر (أي التعلم من المصنف هو إشرافي بما أنه معروف إلى أي فئة ينتمي كل صفوف التدريب (Training Tuples).

التعلم غير الإشرافي Un – Supervised Learning : وهنا نقوم بتقسيم البيانات إلى مجموعات تحتوى على بيانات متشابهة الخصائص لتمييز بعض الاستثناءات بينها, Clustering, عند دراسة الحالة يجب تحديد مجال قاعدة البيانات المستخدمة ، كما يحدد حجم العينة حيث يمكن أن لا نجرى التنقيب على كامل البيانات ويمكن اختيار بعض الصفوف من خلال عينات عشوائية.

يعني التعلم الغير مشرف : إذا كان الـ Class Labels لكل صفوف التدريب Training Tupels غير معروف , وعدد أو مجموعة الـ Classes التي يمكن تعلمها قد لا تكون معروفة مقدما.

2-3-5تقسيم البيانات :

تنقيب البيانات في محتوى ضخم من البيانات غير عملي ويأخذ الكثير من الوقت تقسيم البيانات مفيد لكسب مجموعه من البيانات بدون حدوث فقد في البيانات وجعلها متكاملة حيث يتم تقسيم السجلات بناء على اختبار الخاصية التي تعطي أفضل , ويجب تحديد الأتي :

أولا : تحديد الكيفية التي يتم بها تقسيم السجلات : حيث يتم تحديد شرط اختبار الخاصية ويعتمد على نوع الخاصية (اسمية , مرتبة , مستمرة) , ويعتمد أيضا على طريقة التقسيم (تقسيم ثنائي , متعدد).

تحديد أفضل تقسيم باستخدام معايير اختيار أفضل تقسيم وهي معيار الحصول على المعلومات ومعدل الكسب ومؤشر جيني.

ثانيا : تحديد زمن توقف التقسيم : نتوقف عندما كل السجلات تعود لنفس الصنف , نتوقف عندما تكون كل السجلات لها نفس قيمة الخاصية المماثلة , الإنهاء المبكر[10]

2-3-6مقاييس اختيار الصفة :

هي إرشادات لاختيار معيار التقسيم الأفضل الذي يقسم بيانات معينة D , مقاييس اختيار السمة تعرف أيضا باسم قواعد التقسيم لأنها تحدد كيفية تقسيم السجلات في عقدة معينة.

مقاييس اختيار الصفة توفر تصنيف لكل صفة تصف الصف المعين , الصفة التي لديها أفضل مجموع علامات يتم اختيارها كصفة لتقسيم الصف المعين توجد ثلاثة معايير عامة لاختيار الصفة المناسبة للتقسيم الأفضل وهي :

1. **الحصول على المعلومات information gain**

يعرف اكتساب المعلومة بالانخفاض المتوقع للمعلومات اللازمة لتصنيف الصفة A , الصفة التي تكون قيمة اكتساب المعلومة لها أعلى من قيمة الصفة الأخرى يتم اختيارها كصفة تقسيم للعقدة N في مجموعة البيانات D.

ولحساب اكتساب المعلومة لصفة منفصلة A في مجموعة البيانات D نحسب الأتي :

أ. المعلومات المتوقعة اللازمة لتصنيف سجل في D أو ما يسمى Entropy(D) من المعادلة التالية :

$$\textbf{Info (D)} = \sum_{i=1}^{m} pi \log_2(pi)i \quad \dots(1)$$

$$\textbf{Entropy (D)} = \sum_{i=1}^{m} pi \log_2(pi)i$$

$$Pi = \frac{|Ci,D|}{|D|}$$

حيث Ci,D عدد السجلات في D المنتمية إلى الـ class C , D: العدد الكلي للسجلات.

تستخدم دالة اللوغريثم بالأساس 2 لأن المعلومات ترمز بالـ bits.

- Info(D) هي متوسط كمية المعلومات اللازمة لتحديد فئة السجل i في D.

ب. نحسب المعلومات المتوقعة لـ attribute A في D حيث A تحتوي على عدد V من القيم المنفصلة حيث V={a1,a2,…,av} من المعادلة التالية:

$$info_A\ (D) = \sum_{j=1}^{v} \frac{|Dj|}{|D|} * info(Dj) \quad \dots(2)$$

أو

$$Entropy_A\ (D) = \sum_{j=1}^{v} \frac{|Dj|}{|D|} * Entropy(Dj)$$

حيث $\frac{|Dj|}{|D|}$ وزن الجزء J بالنسبة لقيم الصفة A , Dj عدد الـ tuples التي تنتمي لـ class j التابعة لـ attribute A في D.

ت. لحساب معادلة Information Gain(A) نحسب الفرق بين المعادلة (1) والمعادلة (2)

$$Gain(A) = info(D) - info_A\ (D) \quad \dots(3)$$

أو

$$Gain(A) = Entropy(D) - Entropy_A\ (D)$$

التقسيم بواسطة معيار information gain يعتمد على اختبار عدد المخرجات أي أن هذا المعيار يفضل اختيار الـ attribute ذات عدد أكبر من القيم , فمثلا الصفات ذات القيم الوحيدة unique attribute (مثلا product_id) سوف تكون محل اهتمام لهذا المعيار وبالتالي فإن التقسيمات الناتجة

عنه سوف تكون كبيرة جدا ووحيدة pure (بعدد قيم الـ unique attribute) ولذلك فإن information_attribute(D)= , وبالتالي فإن هذا المعيار غير مناسب لعملية التصنيف لأنه ينتج عددا كبيرا جدا من التقسيمات.

2. معدل الكسب gain ratio

نسبة الكسب امتدادا للكسب المعلوماتي يتخلص من نقاط ضعفه ويطبق نوعا من التطبيع على الكسب المعلوماتي باستخدام معلومات التقسيم Split Information لقيمة معرفة بشكل قياسي كما في المعادلة الآتية :

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|Dj|}{|D|} * \log_2\left(\frac{|Dj|}{|D|}\right).$$

هذا المعيار يأخذ في الاعتبار حساب عدد المخرجات الناتجة لاختبار الصفة , معيار تقسيم لتحديد الانقسام الأفضل يعرف بالمعادلة الآتية :

$$GainRatio\,(A) = \frac{Gain\,(A)}{SplitInfo_A\,(D)}$$

الصفة التي يكون عندها الكسب أعلى قيمة يتم اختيارها كأفضل صفة تقسيم للمجموعة D.

3. مؤشر جيني gini index

معيار مؤشر جيني يحاول تجنب مشكلة الكسب المعلوماتي Information Gain بتقييد شروط الاختبار إلى تقسيمات ثنائية فقط بمعنى يتم حساب مجموع الشوائب الموزونة لكل قسم , ويستخدم هذا المعيار مع خوارزمية CART.

Gini Index يقيس شوائب D ومجموعة سجلات التدريب أو أقسام البيانات حسب المعادلة التالية :

$$Gini\,(D) = 1 - \sum_{i=1}^{m} pi^2$$

حيث $pi = |C_{i,D|}/\,|D|$

الصفة التي لها قيمة أقل من قيمة الصفات الأخرى يتم اختيارها كصفة تقسيم لمجموعة بيانات التدريب D.

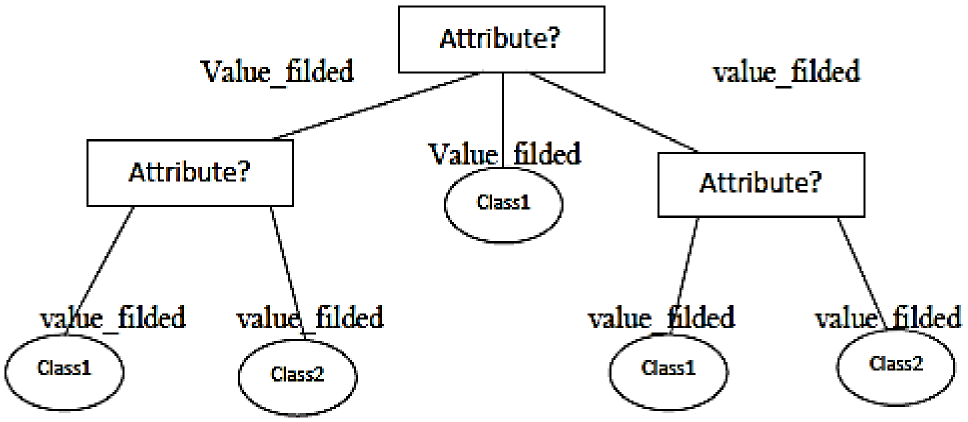
7-3-2 تقنيات التصنيف :

7-3-2-1تقنية أشجار القرار Decision Trees :

وهي نموذج استكشافي يظهر على شكل شجرة كما يعبر اسمها وبشكل دقيق يمثل كل فرع من فروعها سؤالا تصنيفيا وتمثل أوراقها أجزاء من قاعدة بيانات تنتمي للتصنيفات التي تم بنائها .
تتكون شجرة القرارات من سلسلة من العقد والأفرع، ثم تتفرع العقد إلى عقد فرعية بواسطة الأفرع، حيث تمثل العقد السمات التي يجب اعتبارها في عملية صنع القرار ، ثم تأتي القيم المختلفة للسمات من الأفرع المختلفة , وباستخدام نموذج شجرة القرارات في صنع القرارات يمكن البحث من الجذر إلى الأوراق , فتحتوي عقد الأوراق على نتائج كل تصنيف.

1. مصنف شجرة القرار Decision Tree classifier :

أشجار القرار تصنف الحالات عن طريق فرزها على أساس قيم الصفة attribute , كل عقدة داخلية في شجرة القرار تمثل صفة اختبار test attribute , كل فرع يمثل قيمة العقدة , وكل عقدة طرفية (ورقة) تمثل مخرجات الاختبار (قيمة value) , أعلى عقدة في الشجرة هي جذر الشجرة , وتصنف الحالات انطلاقا من عقدة الجذر في شجرة القرار , يتم تعيين كل عقدة ورقية كفئة class.



شكل رقم (7) يوضح شجرة قرار التصنيف

2. استخدام شجرة القرار في التنبؤ :

بالرغم من أن شجرة القرار تستخدم في الاستكشاف وتحضير البيانات للعمليات الإحصائية إلا أنها أيضاً تستخدم وبشكل أكثر للتنبؤ. ومن المهم جداً عند بناء خوارزمية شجرة القرار أن يؤخذ بعين الاعتبار أن تكون قابلة للتطبيق بقدر الإمكان وبشكل مثالي على كل البيانات المتوفرة.

القاعدة الأساسية في بناء شجرة القرار هي إيجاد أفضل سؤال عند كل فرع من فروع الشجرة بحيث يقسم هذا السؤال البيانات إلى قسمين، القسم الأول منها ينطبق عليهم السؤال والقسم الثاني لا ينطبق، وهكذا يتم من خلال سلسلة من الأسئلة بناء شجرة القرار بفروعها المتسلسلة.

2-7-3-2 خصائص تقنية شجرة القرار :

1. مفيدة في التنبؤ والتصنيف.
2. تقسم السجلات إلى أصناف.
3. توقع سلوك الفرد عن طريق مقارنة سلوك الأفراد في المزايا.
4. جيدة التعامل مع مجموعة كبيرة من المعلومات.
5. تحاكي البشر في كيفية اتخاذ القرارات.

2-3-7-3مميزات شجرة القرارات :

1. غير مكلفة للبناء
2. سريعة جدا في تصنيف السجلات الغير معروفه
3. سهلة التفسير للأشجار صغيرة الحجم
4. سهلة الفهم
5. الدقة جيدة , ولكن تُوقف نجاح استخدام شجرة القرار على البيانات التي نعالجها.

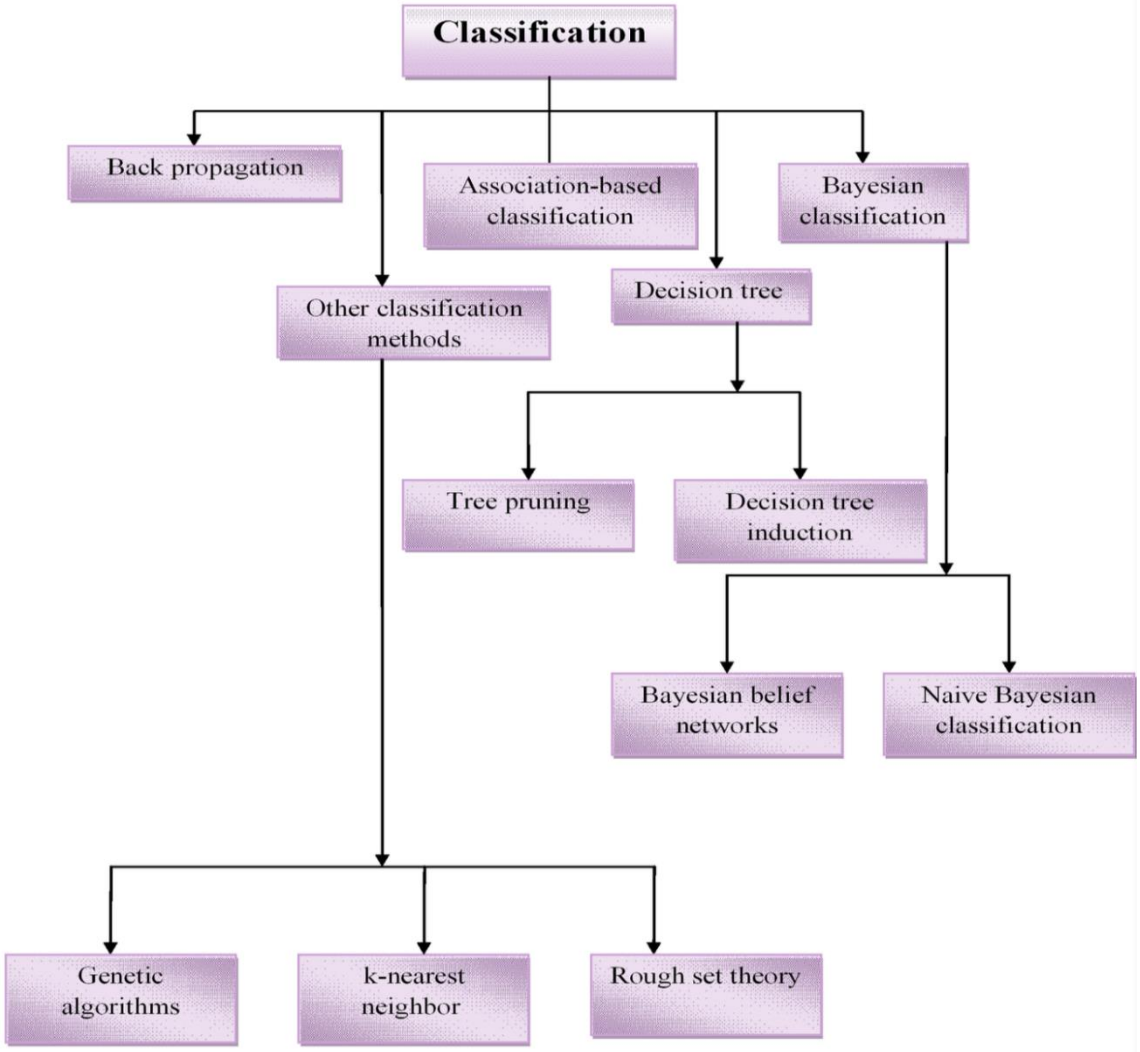
2-3-7-4عيوب شجرة القرارات :

1. يمكن أن تعاني من زيادة المقياس
2. التصنيف يكون بالمستطيلات المقسمة
3. ليست سهلة للبيانات غير العددية
4. يمكن أن تكون ضخمة في حالة ضرورة التجزئة[11].

2-3-7-5 Visualization :

نعني به كيف نستطيع أن ننقل البيانات إلى المستخدم بشكل مفيد ويهدف إبصار البيانات إلى التواصل مع البيانات بشكل واضح وفعال حيث يوفر أدوات فعالة لفهم البيانات , ويعرض الأشكال البيانية , ويمكننا الاستفادة من إبصار البيانات في اكتشاف العلاقة بين البيانات , حيث أنه من

الصعب إيجاد هذه العلاقة بمجرد النظر إلى البيانات الخام , أصبح إبصار البيانات يستخدم بكثرة في الكثير من التطبيقات , على سبيل المثال يستخدم في الأعمال لإصدار التقارير , وإدارة الأعمال , ومتابعة تنفيذ المهام , إن مرئية النتائج تكون مساعدًا في تسهيل ملاحظات مخرجات خوارزميات تنقيب البيانات وفهمها.



شكل رقم (8) التصنيف

الشكل يوضح تقنيات (خوارزميات)التصنيف وهو شكل من أشكال تحليل البيانات التي تستخرج النماذج التي تصف طبقات البيانات الهامة , تعرف هذه النماذج بالمصنفات

DATA MARTS 4-2 (سوق البيانات)

2-4-1 ما هو data mart(سوق البيانات)

إن سوق البيانات data marts هو شكل او نموزج بسيط من مستودع البيانات يركز على موضوع واحد (أو مجال وظيفي) مثل المبيعات أو المالية أو التسويق .وغالبا ما يتم بناء data marts والتحكم بها من قبل إدارة واحدة داخل المنظمة .وبالنظر إلى تركيزها على موضوع واحد، فإن البيانات عادة ما تستمد البيانات من مصادر قليلة فقط ويمكن أن تكون المصادر أنظمة تشغيل داخلية أو مستودع بيانات مركزي أو بيانات خارجية.

2-4-2الاختلافات الأساسية بين مستودع البيانات وسوق البيانات

وعلى النقيض من ذلك، فإن مستودع البيانات يتناول مجالات موضوعية متعددة، وينفذ عادة ويتحكم فيه وحدة تنظيمية مركزية مثل مجموعة تكنولوجيا المعلومات المؤسسية IT group في كثير من الأحيان، ويسمى مستودع البيانات المركزية [12]

يلخص الجدول التالي الاختلافات الأساسية بين مستودع البيانات وسوق البيانات

جدول رقم (4) المقارنة بين data mart&data warehouse

	Data Warehouse	**Data Mart**
Scope	Corporate	Line-of-Business (LoB)
Subjects	Multiple	Single Subject
Data Sources	Many	Few
Size (typical)	100 GB-TB+	< 100GB
Implementation Time	Months to years	Months

2-4-3الإدارة الذكية والتسويق

2-4-4الادارة الذكية:

الإدارة الذكية للأعمال تشير إلى التطبيقات التي تستخدم لجمع البيانات والمعلومات عن عمليات المؤسسة وتحليلها .وتساعد الإدارة الذكية في معرفة العوامل التي تؤثر على أنشطتها التجارية وبالتالي تساعد الشركات على اتخاذ قرارات أفضل . الإدارة الذكية للأعمال تمكن المؤسسات من التنافس بفعالية أكبر.

تساعد الإدارة الذكية المؤسسات على تحليل الاتجاهات المتغيرة في حصص السوق، التغير في سلوك المستهلك، أنماط الإنفاق، وتفضيلات المستهلكين.

الإدارة الذكية تساعد المؤسسة في التخطيط الاستراتيجي وذلك اعتمادا على مفهوم أساسي وهو أن كل نشاط بالمؤسسة يجب أن يخدم شخص ما بشكل مباشر أو غير مباشر وهذا يعني أن القائمين على العمل في حاجة إلى معرفة من هم المستفيدون وماذا يحتاجون لتلبية رغباتهم.

التسويق:

التسويق هو مجموعة من و الأنشطة التي تقوم بإكتشاف رغبات العملاء وإشباعها وتحقق للمؤسسة الربحية خلال فترة مناسبة.

من الناحية المجتمعية، التسويق هو الرابط بين احتياجات المجتمع وبين الاستجابة لأنماط الاقتصاد من خلال توصيل قيمة منتج أو خدمة إلى العملاء .أما التسويق كعلم، فهي عملية تحديد السوق المستهدف من خلال القيام بتحليل السوق وفهم توجهات العملاء وتقديم قيمة عالية لهم.

يتألف النشاط التسويقي من أربعة عناصر أساسية حددها" جيروم ماكارتي -:"

1-المنتج:

يعني مايتم عرضه في السوق ، الخدمات التي يتحصل عليها المشتري.

2-التوزيع او المكان:

الطريقة التي تجعل المنتج في متناول المشتري ووصوله إلى السوق المستهدف.

3-الترويج:

يعني نشاطات الاتصال مثل الإعلان عن المنتج و السوق المستهدف المتواجد به المنتج.

4-السعر:

يعني سعر المنتج بالإضافة إلى الرسوم الأخرى مقابل التوصيل والتصريح

ثم جاء عملاق التسويق كوتلر وأضاف عنصرين أساسيين آخرين وهما:

1-السياسات:

السياسات المتبعة في الدولة تؤثر على التسويق وبالتالي الإنتاجية.

2-الراي العام:

راي المجتمع يؤثر على إنتاجية المنتج لذلك لابد أن يكون محبب له.

استراتجيات التخطيط للتسويق الذكي:

رسالة الشركة : لابد من وجود رسالة واضحة للشركة ليعرف الناس عن شركتك .

الاهداف الرئيسية : وجود اهداف رئيسية تساعد صاحب راس المال على التركيز في عمله وابتكار خطوات وطرق من اجل تحقيق هذه الاهداف الرئيسية.

المراجعة التسويقية : مراجعة السوق والطريقة التسويقية للمنتجات او الخدمة التي يقدمها صاحب راس المال ، وهل هناك تقبل لطريقة التسويق التي يستخدمها صاحب المنتج او الخدمة

تحليل الشركة والبيئة والمستهلك: دراسة السوق والبيئة والمستهلك لتجنب اي خسائر او مشاكل قد تواجه الشركة مستقبلا.

وضع الفرضيات لمواجهة اي تحدي : وضع حلول افتراضية مستقبلية لمواجهة اي مشاكل تحديات لبعض المنافسين او مشاكل حروب او اضطرابات سياسية واخرى.

الاهداف التسويقية والاستراتجيات : طريقة وضع الاهداف التسويقية والخطط التسويقية والاستراتجيات المستخدمة من فترة الي فترة معينة على شكل تسلسل زمني وتسلسل من

الانجازات التي تخدم اهداف الشركة.

تقدير النتائج المتوقعة : وضع نتائج متوقعة بين كل فترة زمنية معينة لقياس النتائج والمشاكل التي تواجه الشركة او المشروع.

وضع الخطط البديلة والطوارىء: الخطط البديلة التي قد تطرأ بسبب الحروب او مشاكل سياسية او اقتصادية.

الموازنة : وضع ميزانية مناسبة للتخطيط وهي المحرك الوحيد لكل الخطط وهي احدى الاسباب الرئيسية لنجاح المشاريع وهوالقلب الذي يبث الحياه في الشركة والمشاريع الخ .

2-4-5التسويق الذكي واستهداف العملاء

يعتمد التسويق الذكي Smart marketing) ويسمى اختصاراً (smarketing على الاختيار الأمثل والصحيح لمجموعة الزبائن المراد استهدافهم بالخطة التسويقية ، وكلمة الاختيار الذكي للذبون ترتبط ارتباطاً وثيقاً بمعرفة الزبون ، بدءاً من أبسط الأمور ، كمكان سكن مجموعة الزبائن المستهدفين ، وصولاً إلى معلومات أكثر دقة : (أين يعمل . هل يملك أطفال – ما المجلات التي يقرأها – ما البرامج التلفزيونية التي يتابعها – ما الأمور التي يكرهها – ما الأمور المميزة فيه) خلاصة القول أن التسويق الذكي يعتمد على بنك المعلومات الذي لديك حول الزبائن المستهدفين.

حالما تتوفر هذه المعلومات بين يديك ، يمكنك البدء باستخدمها لعرض منتجاتتك أون لاين ، عملية اختيار المنتج ثم اختيار الزبون الذي هو بحاجة هذا المنتج و عرضه أمامه تسمى " استهداف الزبون " ، ولكن كيف ستقوم بتحليل هذا الكم الكبير من البيانات لاستخلاص المجموعة المناسبة من الزبائن و التي من المحتمل أن تقوم بشراء المنتج ؟

في الحقيقة يوجد كم كبير من البرمجيات و الأدوات المختصة في هذا المجال بدءاً من عملية جمع المعلومات ، إلى تحليلها ، إلى اختيار مجموعة الزبائن المناسبة ، وصولاً إلى عرض الإعلان أمام الزبون المرغوب وذلك فيما يعرف بتنقيب البيانات او الـ data mining

القواعد الذهبية في التسويق الذكي:

أ. القاعدة الأولى : اجمع معلوماتك باختيار الأسئلة الأنسب:

الأمر الذي يميز التسويق الذكي عن غيره هو " الاستهداف الذكي للزبائن " ، التسويق الذكي يهدف إلى بناء علاقة ثقة بين المنتج و الزبون المستهدف بالعملية التسويقية فاختيار الزبون يعتمد بشكل أساسي على مدى الخبرة في جمع المعلومات.

من الأمور التي يمكن أن تساعدك في عملية جمع المعلومات هي دراسة الشركات المنافسة ، ومعرفة الأماكن التي نجحوا فيها و الأماكن التي فشلوا فيها ، ومراقبة نوعية المعلومات التي يقومون بجمعها ، هذه الدراسة و المراقبة للمنافس تساعدك كثيراً في جمع صحيح للمعلومات ، وبالتالي تسويق ذكي ناجح و مؤثر.

ب. القاعدة الثانية : استخدم معلوماتك بالشكل الأنسب:

بعد أن انتهيت من جمع المعلومات حان الوقت لتوظيف هذه المعلومات و استخدامها في خطتك التسويقة ، في الحقيقة هناك الكثير من المواقع و الأدوات المتاحة على الأنترنت و التي يمكنك توظيفها بهدف تحقيق تسويق ذكي للمنتج أو الخدمة ، هذه الأدوات تقدم لك الكثير من المعلومات المفيدة في مجال التسويق الذكي ، فهي تساعدك في تحديد أين يقضي زبائنك المستهدفين أوقاتهم على الانترنت و تقدم لك الكثير من المخططات البيانية و المعلومات حول نشاطات الزبائن على الانترنت ، بعد انتهائك من جمع المعلومات عليك التوجه إلى هذه الأدوات و التدرب عليها و التأقلم معها لأنها ستصبح شريكك الأساسي في العمل لاحقاً

الآن وبعد أن قمت بجمع معلومات الزبائن و معرفة احجام متطلباتهم ، حان الوقت للبدء باستهداف الزبائن و البدء بالتسويق بعد تنقيب تلك البيانات البيانات ، هناك الكثير من الأماكن و المواقع التي يمكن استغلالها للتسويق الإلكتروني

الشبكات الإجتماعية :

الشبكات الإجتماعية هي من الوسائل الفعالة في الإعلان الإلكتروني ، فهي تضم كم هائل من المستخدمين ، و تلقى رواجاً كبيراً مقارنة بغيرها من المواقع

التسويق بواسطة الهواتف الذكية

نلاحظ كثيراً تأثير الهواتف الذكية على حياتنا كمستهلكين وكيف أنها أصبحت تمثل ضرورة قصوى في حياتنا , ليس فقط للتواصل مع الأصدقاء ولكن لممارسة كل شيء في حياتنا , اللعب والتسلية , القراءة , الكتابة , التسوق , التواصل الاجتماعي, العمل , وغيرها من الأمور الضرورية في أهم أدوات التسويق عن طريق الهواتف الذكية هي:

1. التسويق باستخدام الرسائل القصيرة وتطبيقات المحادثة:
2. whatsapp
3. تطبيقات الجوال Mobile App:
4. الاعلانات ومحركات البحث Mobile Advertising and paid search:

ج.القاعدة الثالثة : إفهم زبائنك جيداً:

إذا لم تنجح سياسة التسويق عبر مواقع التواصل الإجتماعي و لم تأت بالنتائج المرجوة ، لا تقلق فهذه ليست النهاية . هناك عدد لا نهائي من المواقع و الشبكات الأساسية التي يمكنك توظيفها في التسويق الذكي ، النقطة الجوهرية هي فهم الزبائن المستهدفين ، ومعرفة أين يمضون أوقاتهم على شبكة

الإنترنت ، فكلما كنت ناجحاً في فهم زبونك كلما كان تسويقك أكثر ذكاءاً ، لفهم زبائنك بالشكل الأمثل هناك الكثير من الأدوات المتوفرة على الانترنت و التي تساعدك في هذا الموضوع وتمكنك من معرفة أين يقضي زبائنك أوقاتهم.

بالنهاية فإن الجمع بين القواعد الثلاثة التي ذكرتها يضمن لك النجاح ، فجمع المعلومات الصحيح و حسن تحليلها و توظيفها ، و المهارة في فهم الزبون ، بالإضافة إلى براعتك في استخدام الأدوات الخاصة بالتسويق سيجعل منك محترفاً في مجال التسويق الإلكتروني الذكي[13].

5-2 برنامج **Weka**

WEKA عبارة عن حزمة برمجية مفتوحة المصدر تحتوي على مجموعة من الخوارزميات التي تساعد على تحليل وإستخراج البيانات (Data mining) هذه الخوارزميات يمكن تطبيقها بسهولة على مجموعة من البيانات إما بشكل مباشر من خلال واجهة برنامج WEKA أو من خلال استدعاؤها عن طريق (Java code) بإستخدام classes الخاصة بها, وذلك من خلال تحميل المكتبة الخاصة بـWEKA

تعد الأداة مفتوحة المصدر Weka من أشهر أدوات برمجيات التنقيب عن البيانات ، والتي تدخل في العديد من الأدوات الأخرى مثل KNIME.، Rattle، RapidMiner .

WEKA هو برنامج خاص بخوارزميات التنقيب عن البيانات وتعلم الآلة , تمت كتابة هذا البرنامج من قبل جامعة Waikato في نيوزيلاند وهو مكتوب بلغة الجافا , يعمل تحت رخصةGPL , وهو يعمل تحت عدة بيئات (لينكس , ويندوز , ماك).

WEKA هو عبارة عن حزمة برمجية مفتوحة المصدر تحتوي على مجموعة من الخوارزميات التي تساعد على تحليل واستخراج البيانات Data Mining هذه الخوارزميات يمكن تطبيقها بسهولة على مجموعة من البيانات من خلال واجهة برنامج WEKA الواجهة الرسومية التي تتكون من خمسة تطبيقات رئيسية، وهي

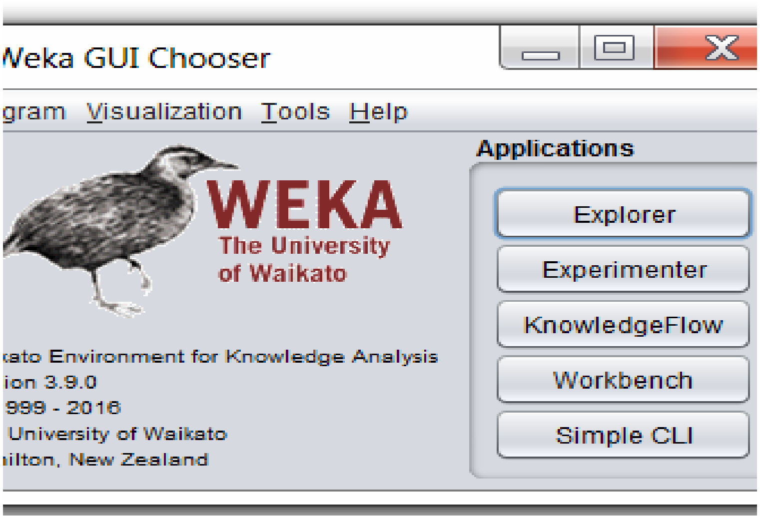
1. المستكشفExplorer لاستكشاف البيانات
2. المجربExperimenter للقيام بالتجارب والاختبارات الإحصائية
3. وتدفق المعرفةKnowledgeflow
4. وواجهة الأوامر Workbench لتنفيذ أوامر الويكا
5. امكانية الاضافة والبرمجةSimple CLI [14]

يحتوي هذا البرنامج على أدوات قادرة على التعامل مع الأمور التالية :

, Classification , Regression , Clustering , Association Rules Pre-processing , and Visualization.

الخصائص الرئيسية لـ WEKA تحتوي على:

1. أدوات المعالجة الأولية , خوارزميات التعلم وطرق التقييم.
2. واجهة المستخدم الرسومية (data visualization).
3. بيئة لمقارنة خوارزميات التعلم.



شكل رقم(9) واجهة الاداة مفتوحة المصدر WEKA

عند فتح الملف المراد تصنيفه في Weka يتم اختبار تصنيفة حسب الطرق المذكورة ادناه وتقييم المصنف

تقييم المصنف Evaluating

هو إجراء اختبار على المصنف وذلك بإعطائه مجموعة من البيانات ليصنفها حسب ماتعلم مسبقًا.

اختبار المصنف

اختبار المصنف مهم جدًا لنعرف هل المصنف الذي تم تدريبه يمكن استخدامه في تصنيف بيانات جديدة أم لا, نحدد ذلك من خلال نسبة تصنيفه في مرحلة الإختبار, كلما كانت عالية كانت امكانية استخدامه أكبر.

طرق اختبار المصنف

هناك أربعة طرق مستخدمة في weka تستخدم في اختبار المصنف:

1. Using Training set
2. Cross validation
3. Percentage Split
4. Supplied test set

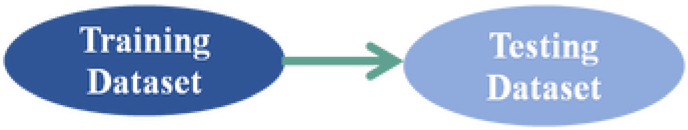
سنتعرف الآن على كل من هذه الطرق وكيفية استخدامها!..

1. Using Training set

هذه الطريقة من اسمها يتضح لنا أن البيانات التي نستخدمها في اختبار المصنف هي مجموعة التدريب التي تم تدريبه عليها.

كيف يكون ذلك؟ يقوم المصنف بأخذ مجموعة التدريب التي تدرب عليها وتغطية (label) الذي تم اعطائه للبيانات ويبدأ المصنف بتخمين label لها.

-> labelهو التصنيف الذي يعطي البيانات مثل.positive , negative



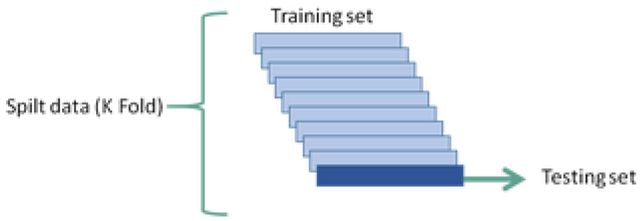
2. Cross validation

يتم استخدام البيانات التي تدرب عليها مسبقًا كما في النوع الأول, لكن ماهو الفرق بينهم؟

في هذه الطريقة يتم تقسيم البيانات حسب الرقم الذي يعطى (fold) ومن ثم يتدرب على 90% من هذه البيانات المقسمة و 10% يتم اختبار المصنف عليها.

مثال: لدينا مجموعة تدريب عددها 100 استخدمنا cross validation في اختبار المصنف, تُرى ماذا يحصل لهذه البيانات؟

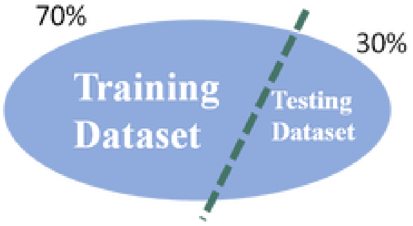
أولاً يتم تقسيم البيانات حسب (fold) مثلاً لوأعطيناه قيمة 10, يتم تقسيم البيانات إلى 10 مجموعات كما في الصورة أدناه.



ثم يتم تدريب المصنف على تسع مجموعات والعاشرة يتم اختباره عليها, ومن ثم يتم إجراء الإختبار على مجموعة أخرى وإدخال المجموعة السابقة التي أجري عليها الاختبار ضمن مجموعة التدريب وهكذا حتى يتم اختبار المصنف عشر مرات .هذه الطريقة مناسبة عندما تكون مجموعة البيانات قليلة.(

3. Percentage Split

في هذه الطريقة يتم تقسيم مجموعة التدريب إلى قسمين مجموعة يتم تدريبة عليها ومجموعة يتم اختباره عليها



4. Supplied test set

في هذه الطريقة يتم ادخال بيانات جديدة لم يراها المصنف من قبل ويتم اختباره عليها.

Classifier Accuracy

تفاصيل نتائج الاختبار للمصنف الذي تم تدريبه مسبقًا واختباره ، سنستخدم ثلاث دوال من evaluate classلإظهار النتائج.

مخرجات هذه الدوال تظهر كما نراها في الصورة.

```
Correctly Classified Instances         194               77.6    %
Incorrectly Classified Instances        56               22.4    %
Kappa statistic                          0.544
Mean absolute error                      0.2476
Root mean squared error                  0.3531
Relative absolute error                 51.5556 %
Root relative squared error             72.0665 %
Total Number of Instances              250

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.767     0.21       0.846     0.767     0.804      0.885    positive
                 0.79      0.233      0.693     0.79      0.738      0.885    negative
Weighted Avg.    0.776     0.219      0.785     0.776     0.778      0.885

=== Confusion Matrix ===

   a   b   <-- classified as
 115  35 |   a = positive
  21  79 |   b = negative
```

شكل رقم (10) عملية تصنيف باستخدام weka

لو لاحظنا في الجزء العلوي من النتائج, بالتحديد في أول سطرين يوجد نسبة التصنيف الصحيح والخاطئ

```
Correctly Classified Instances        194          77.6   %
Incorrectly Classified Instances       56          22.4   %
```

شكل رقم (11) نسبة الحالات الصحيحة في المائة ونسبة الحالات الخطا في المائة

في السطر الأول يبين لنا مجموع الحالات (instances) التي صنفها المصنف بشكل صحيح ونسبتها من مئة. في السطر الثاني يبين لنا مجموع الحالات (instances) التي أخطأ المصنف في تصنيفها.

لو نظرنا في أسفل النتائج لوجدنا مصفوفة تحوي على أرقام مختلفة.

مصفوفة الخطأ (Confusion Matrix)

```
=== Confusion Matrix ===

   a   b   <-- classified as
 115  35 |   a = positive
  21  79 |   b = negative
```

شكل رقم (12) يوصف مصفوفة التضارب

هذه المصفوفة تسمى مصفوفة الخطأ أو :(Confusion Matrix) وهي مصفوفة مشهورة في مجال تعليم الآلة, تستخدم لاختبار أداء الخوارزميات. وهي عبارة عن جدول يحوي على معلومات وتفاصيل حول التصنيفات الفعلية (التي تم تصنيفها من قبل الإنسان) والتصنيفات التنبئية التي تنبئها المصنف. فكل عامود في المصفوفة يمثل الفئات المُتنبئة (Predicted class) و كل صف يمثل الفئات الفعلية (Actual class).ويتم تقييم أداء المصنف عادةً باستخدام البيانات الواردة في المصفوفة, ويكون حجم المصفوفة حسب عدد الفئات. يبين الجدول التالي شكل المصفوفة ومعنى كل عامود وصف فيها وكل خانة في المصفوفة التي سنشرح الأن معناها.

	p' (Predicted)	n' (Predicted)
p (Actual)	True Positive	False Negative
n (Actual)	False Positive	True Negative

شكل رقم (13) الفئات المُتنبئة و الفئات الفعلية

P,N يمثلون الفئات الموجودة. كل إدخال (خانة) في هذه المصفوفة له معنى وسنوضحه الآن:

1. TP)True Positive): عدد الحالات التي تنبئها المصنف بشكل صحيح لفئة p و تنتمي بشكل فعلي لفئةp
2. FN) False Negative): عدد الحالات التي تنبئها المصنف بشكل غير صحيح لفئة p و تنتمي بشكل فعلي لفئةn
3. TN) True Negative): عدد الحالات التي تنبئها المصنف بشكل صحيح لفئة n و تنتمي بشكل فعلي لفئةp
4. FP) False Positive): عدد الحالات التي تنبئها المصنف بشكل غير صحيح لفئة n و تنتمي بشكل فعلي لفئةp

دعونا ننتقل للجزء الأوسط من المخرجات, الذي يرتبط بمصفوفة الخطأ.

```
=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall   F-Measure   ROC Area   Class
                   0.767     0.21      0.846      0.767     0.804      0.885     positive
                   0.79      0.233     0.693      0.79      0.738      0.885     negative
Weighted Avg.      0.776     0.219     0.785      0.776     0.778      0.885
```

شكل رقم (14) نسبة الصحة والخطأ والدقة

نفهم من الملاحظة أعلى الصورة أن هذا الجزء تفاصيل لنسبة الصحة والخطأ والدقة وكما نلاحظ أن هذه المسميات حسبت لكل فئة لدينا مثلاً في الصورة أعلاه حسبت لـ فئتي positive وnegative دعونا الآن نتعرف على تفاصيل هذه الجزئية وماذا يعني كل مسمى لدينا في هذه الجزئية؟ وكيف سنستخدم المصفوفة؟

- **(Recall or True positive rate (TP Rate**

معدل الحالات التي صنفت لفئة معينة.

مثلاً: في الصورة أعلاه معدل الحالات التي أعطاها المصنف تصنيف 767 positive=,وتحسب كما يلي:

TP/ TP+FN

- **(False positive rate (FP Rate**

معدل الحالات التي صنفت بشكل غير صحيح لفئة معينة. وتحسب كما يلي:

FN / TP+FN

- **Precision**

معدل الحالات الصحيحة لفئة معينة, وتنبئها المصنف بشكل صحيح.

مثلاُ: لدينا فئة positive معدل التي تنبئها المصنف على إنها فئة positive بشكل صحيح =0.767. وتحسب كما يلي: [15]

TP/ TP+FP

2-6 نبذة عن شركة سويتش للتجارة والهندسة:

تأسست عام 1988 كشركة خاصة ذات مسئولية محدودة بقانون الشركات لعام 1925م تحت اسم سويتش للهندسة الكهربائية المحدودة ,ومنذ ذلك الحين وهي في حالة نمو مضطرد وتنوع مستمر في مختلف مجالات انشطة القطاع الخاص,وفي العام 2007 تحولت الى مجموعة من الشركات المستقلة تعمل تحت مظلتها (شركة كاف للطرق والجسور,شركة هيلث كير الطبية,شركة سويتش للخدمات الزراعية .شركة ايديا لخدمات ما بعد البيع وشركة سويتش القابضة)هذه الشركات قامت على اساس قوي واضح المعالم والمبادئ التجارية

اقدم نشاط تجاري في شركة سويتش هو بيع المولدات وهي من اوائل الشركات صاحبة التوكيل في هذا المجال وقد كانت وما زالت الوكيل الحصري لمولدات SDMO التي كانت فرنسية ثم اصبحت امريكية لم تقف المقاطعة الامريكية في استمرار الوكالة لان الشركة كانت من العملاء الاساسيين لدا SDMO تتميز هذه الماركة بأنها صاحبة خبرة اكثر من 40 سنة في مجال توليد الطاقة والاولى على المستوى الاوربي ومتوفرة باحجام من 0.5-KVA3300 (KVA=1000 VA) , لشركة سويتش توكيلات عديدة مثلا تمتلك توكيل UPS RIELLO الايطالي لبطاريات تخزين الطاقة كما تمتلك توكيل LORENIZ الالماني لطلمبات الطاقة الشمسية وعدة توكيلات مثل شرائح الطاقة الشمسية ،مكيفات,واليات زراعية.[16]